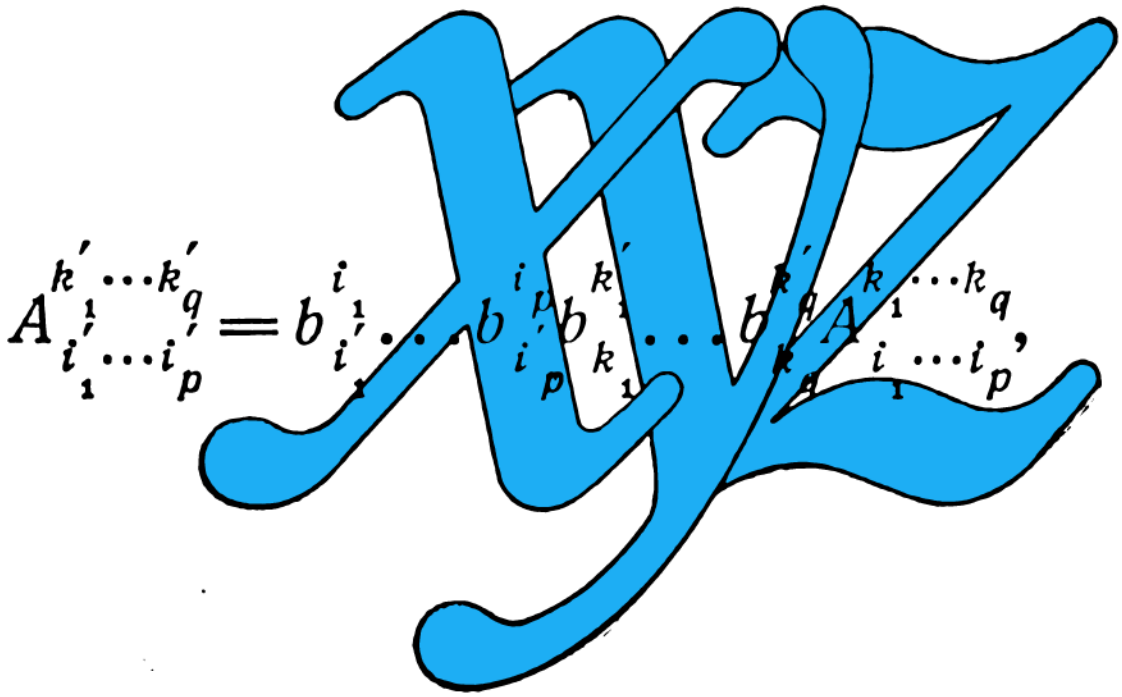
LINEAR ALGEBRA

V.A. ILYIN
and
E.G. POZNYAK

MIR
PUBLISHERS
MOSCOW



This textbook of linear algebra, which is based on a course of lectures read by the authors at Moscow University, is intended for university students specializing in applied mathematics and physics. It discusses matrices and determinants, linear systems of equations, finite dimensional vector and Euclidean spaces, linear functionals and linear operators in these spaces, the theory of linear and quadratic forms, the theory of tensors, classification of second-degree surfaces and problems of group representations. Because of the increased emphasis on computational methods the book presents iteration methods of solving linear systems of equations and corresponding eigenvalue problems as well as A. N. Tychonoff's regularization method as applied to incorrectly set problems of linear algebra.

Professor ILYIN Vladimir Aleksandrovich is Head of a Chair at Moscow State University, senior research worker at the Institute of Mathematics of the USSR Academy of Sciences, Lomonosov Prize and USSR State Prize winner. He is the author of a number of textbooks for universities and more than a hundred scientific papers on the theory of partial differential equations, on the theory of functions and functional analysis and on the spectral theory of differential operators.

His most important scientific achievements concern the establishment of exact conditions for the expansion of elliptic operators in eigenfunctions and of exact conditions for convergence of spectral expansions corresponding to nonself-adjoint differential operators, and also the search for conditions for solvability of boundary-value and mixed problems for differential operators with nonsmooth coefficients and nonsmooth boundary manifolds.

Professor POZNYAK Eduard Genrikhovich, D.Sc. (Phys.-Math.) is a lecturer of mathematics at the Physics Department of Moscow State University, USSR State Prize winner. He is the author of a number of university textbooks and more than 70 scientific papers dealing with various problems of geometry and differential equations related to geometry.

His most important results are connected with the theory of isometric embedding in the whole of two-dimensional Riemannian manifolds of negative curvature in Euclidean spaces, investigations connected with Lobachevsky's geometry and with various divisions of natural science.

LINEAR ALGEBRA

V.A. ILYIN
and
E.G. POZNYAK

В. А. Ильин, Э. Г. Позняк

ЛИНЕЙНАЯ АЛГЕБРА

«НАУКА» МОСКВА

V. A. Ilyin and E. G. Poznyak

LINEAR ALGEBRA

Translated from the Russian
by Irene Aleksanova

Mir Publishers Moscow

First published 1986
Revised from the 1984 Russian edition

На английском языке

© Издательство Наука, Главная редакция физико-математической литературы, 1984
© English translation, Mir Publishers, 1986

CONTENTS

Preface 9
Introduction

Chapter 1. Matrices and Determinants 12

1.1. Matrices 12
1.1.1. The concept of a matrix. 1.1.2. The main operations on matrices and their properties. 1.1.3. Hypermatrices
1.2. Determinants 18
1.2.1. The concept of a determinant. 1.2.2. Expressing a determinant directly in terms of its elements. 1.2.3. Laplace's expansion. 1.2.4. Properties of determinants. 1.2.5. Examples of calculations of determinants. 1.2.6. The determinant of the sum and the product of matrices. 1.2.7. The concept of an inverse matrix
1.3. Theorem on the Base Minor of a Matrix 39
1.3.1. The concept of a linear dependence of rows. 1.2.3. Theorem on the base minor. 1.3.3. The necessary and sufficient condition for a determinant to be zero

Chapter 2. Linear Spaces 43

2.1. The Concept of a Linear Space 43
2.1.1. Definition of a linear space. 2.1.2. Some properties of arbitrary linear spaces.
2.2. The Basis and the Dimensionality of a Linear Space 47
2.2.1. The concept of a linear dependence of the elements of a linear space. 2.2.2. The basis and the coordinates. 2.2.3. The dimensionality of a linear space. 2.2.4. The concept of isomorphism of linear spaces
2.3. Subspaces of Linear Spaces 54
2.3.L The concepts of a subspace and a linear closure. 2.3.2. A new definition of a rank of a matrix. 2.3.3. The sum and the intersection of subspaces. 2.3.4. Expansion of a linear space into a direct sum of subspaces
2.4. Transformation of Coordinates in Transformation of the Basis of an n-dimensional Linear Space 61
2.4.1. Direct and inverse transformations of bases. 2.4.2. The relation between a change of a basis and a transformation of the corresponding coordinates

Chapter 3. Systems of Linear Equations 64

3.1. The Condition for the Consistency of a Linear System 64
3.1.1. The concept of a system of linear equations and its solutions. 3.1.2. Non-trivial consistency of a homogeneous system. 3.1.3. The condition of the consistency of a general linear system
3.2. Seeking Solutions of a Linear System 69
3.2.1. A quadratic system of linear equations with a nonzero determinant of the system matrix. 3.2.2. Seeking all the solutions of a general linear system. 3.2.3. The properties of the collection of solutions of a homogeneous system. 3.2.4. Concluding remarks on solution of linear systems

Chapter 4. Euclidean Spaces 82

4.1. A Real Euclidean Space and Its Simplest Properties 82
4.1.1. Definition of a real Eulidean space. 4.1.2. The simplest properties of an arbitrary Euclidean space
4.2. An Orthonormal Basis of a Finite-Dimensional Euclidean Space 89
4.2.1. The concept of an orthonormal basis and its existence. 4.2.2. The properties of an arthonormal basis. 4.2.3. Decomposition of an n-dimensional Euclidean space in a direct sum of its subspace and its orthogonal complement. 4.2.4. The isomorphism of n-dimensional Euclidean spaces
4.3. A Complex Euclidean Space 95
4.3.1. Definition of a complex Euclidean space. 4.3.2. The Cauchy-Bunyakovsky inequality. The concept of a norm. 4.3.3. An orthonormal basis and its properties
4.4. The Regularization Method for Seeking a Normal Solution of a Linear System 99

Chapter 5. Linear Operators 107

5.1. The Concept of Linear Operator. Its Principal Properties 107
5.1.1. Definition of a linear operator. 5.1.2. Actions performed on linear operators. The space of linear operators. 5.1.3. The properties of the set L (V, W) of linear operators
5.2. Matrix Notation for Linear Operators 114
5.2.1. The matrices for linear operators in a given basis of the linear space V. 5.2.2. Transformation of the matrix of a linear operator in passing to a new basis. 5.2.3. A characteristic polynomial of a linear operator
5.3. Eigenvalues and Eigenvectors of Linear Operators 120
5.4. Linear and Sesquilinear Forms in a Euclidean Space 122
5.4.1. A special representation of a linear form in a Euclidean space. 5.4.2. Sesquilinear forms in a Euclidean space. Special representation of such forms
5.5. Linear Self-Adjoint Operators in a Euclidean Space 126
5.5.1. The concept of an adjoint operator. 5.5.2. Self-Adjoint operators. Basic properties. 5.5.3. The norm of a linear operator. 5.5.4. Other properties of self-adjoint operators.
5.5.5. Expansion of self-adjoint operators in terms of eigenfunctions. Hamilton-Cayley theorem. 5.5.6. Positive operators. The mth-degree roots of an operator
5.6. Reducing a Quadatic Form to the Sum of Squares 139
5.7. Unitary and Normal Operators 142
5.8. Canonical Form of Linear Operators 146
5.9. Linear Operators in a Real Euclidean Space 150
5.9.1. General remarks. 5.9.2. Orthogonal operators

Chapter 6. Iterative Methods of Solving Linear Systems and Eigenvalue Problems 158

6.1. Iterative Method of Solving Linear Systems 159
6.1.1. The method of simple iteration (Jacobi's method). 6.1.2. The general implicit method of simple iteration. 6.1.3. A modified method of simple iteration. 6.1.4. Seidel's method. 6.1.5. The method of overrelaxation. 6.1.6. The case of a nonsymmetric matrix A. 6.1.7. Chebyshev's iterative method
6.2. Solving a Complete Problem of Eigenvalues by the Rotation Method 177

Chapter 7. Bilinear and Quadratic Forms 183

7.1. Bilinear Forms 183
7.1.1. The concept of a bilinear form. 7.1.2. Representation of a bilinear form in a finite-dimensional linear space. 7.1.3. Transformation of a matrix of a bilinear form in passing to a new basis. The rank of a bilinear form
7.2. Quadratic Forms 187
7.3. Reducing a Quadratic Form to the Sum of Squares 189
7.3.1. Lagrange's method. 7.3.2. Jacobi's method
7.4. The Law of Inertia of Quadratic Forms. Classification of Quadratic Forms 194
7.4.1. The law of inertia of quadratic forms. 7.4.2. Classification of quadratic forms. 7.4.3. Sylvester's criterion for the definiteness of a quadratic form
7.5. Multilinear Forms 200
7.6. Bilinear and Quadratic Forms in a Euclidean Space 201
7.6.1. Preliminary remarks. 7.6.2. Reducing a quadratic form to the sum of squares in an orthogonal basis. 7.6.3. Simultaneous reduction of two quadratic forms to the sum of squares in a linear space. 7.6.4. Extremal properties of a quadratic form
7.7. Second-Order Hypersurfaces 207
7.7.1. The concept of a second-order hypersurface. 7.7.2. Translations in a Euclidean space. Transformations of orthonormal bases into orthonormal bases. 7.7.3. Transformation of a general equation of a second-order hypersurface in translation. 7.7.4. Transformation of a general equation for a second-order hypersurface when passing from an orthonormal basis to an orthonormal one. 7.7.5. Invariants of the general equation of a second-order hypersurface. 7.7.6. The centre of a second-order hypersurface. 7.7.7. The standard simplification of any equation of a second-order hypersurface by transforming an orthonormal basis. 7.7.8. Simplification of the equation of a central second-order hypersurface. Classification of central hypersurfaces. 7.7.9. Simplification of the equation of a noncentral second-order hypersurface. Classification of noncentral hypersurfaces.

Chapter 8. Tensors 224

8.1. Transformation of Bases and Coordinates Covariant and contravariant vector coordinates. 8.1.1. Gram's determinants. 8.1.2. Reciprocal bases. 8.1.3. Transformation of a basis and coordinates. 224
8.2. The Concept of a Tensor. Principal Operations on Tensors 230
8.2.1. The Concept of a tensor. 8.2.2. Examples of tensors. 8.2.3. The principal operations on tensors
8.3. A metric Tensor. Principal Operations of Vector Algebra in Tensor Designations 239
8.3.1. The Concept of a metric tensor in a Euclidean space. 8.3.2. The

operations of lifting and lowering indices with the aid of a metric tensor. 8.3.3. Orthonormal bases in E^n. 8.3.4. A discriminatory tensor. 8.3.5. An oriented volume. 8.3.6. A vector product. 8.3.7. A triple vector product

8.4. A Metric Tensor of a Pseudo-Euclidean Space 248

8.4.1. The concept of a pseudo-Euclidean space and a metric tensor of a pseudo-Euclidean space. 8.4.2. Galilean coordinates. The Lorentz transformations. 8.4.3. The Lorentz transformations of the space $E^4_{(1,\ 3)}$.

8.5. The Tensor of the Moment of Inertia 254

Chapter 9. Elements of the Group Theory 256

9.1. The Concept of a Group. The Main Properties of Groups 256

9.1.1. Composition laws. 9.1.2. The concept of a group. Some properties of groups. 9.1.3. Isomorphism of groups. Subgroups. 9.1.4. Cosets. Normal subgroups. 9.1.5. Homomorphisms. Factor Groups

9.2. Transformation Groups 267

9.2.1. Nonsingular linear transformations, 9.2.2. A group of linear transformations. 9.2.3. Convergence of elements in the group $GL(n)$. The subgroups of the group $GL(n)$. 9.2.4. A group of orthogonal transformations. 9.2.5. Certain discrete and finite subgroups of an orthogonal group. 9.2.6. The Lorentz group. 9.2.7. Unitary groups

9.3. Group Representations 277

9.3.1. Linear representations of groups. Terminology. 9.3.2. Matrices of linear representations. Equivalent representations. 9.3.3. Reducible and irreducible representations. 9.3.4. Characters. 9.3.5. Examples of group representations

Index

Preface

The textbook is based on the course of lectures delivered by the authors at Moscow University.

There are some peculiarities in the exposition of the material. It begins with the study of matrices and determinants, and the nth-order determinant is introduced by induction by way of the $(n-1)$-order determinant with the aid of the expansion according to the first row. The theorem on the expansion according to any other row and any column is easily proved (the scheme of the proof of that theorem turns out to be quite similar to that of the Laplace theorem). The traditional definition of a determinant in terms of its elements is a simple consequence of the definition given in this book.

The study of linear systems is preceded by the theory of linear spaces and transformations of bases and vector coordinates in those spaces. When studying linear systems we at once introduce to the reader not only the ordinary but also the matrix notation of the system and the derivation of Cramer's formulas.

The study of real and complex Euclidean spaces is concluded with the proof of Tikhonov's theorem on seeking a normal solution of a linear system.

When linear operators are considered, all the main aspects of the spectral theory and finite-dimensional Euclidean spaces are presented. The theorem on the reduction of a matrix to the Jordan form is proved by a brief method, based on induction, suggested by Filippov.

There is a full chapter in the book dedicated to the methods of iteration in which the most important iteration methods of solving linear systems are considered from one and the same point of view (explicit and implicit methods of simple iteration, Seidel's method, the method of overrelaxation), and the conditions for the convergence of these methods are established. The conditions for the fastest convergence, suggested by Samarsky, are ascertained for the general implicit method of simple iteration. The convergence of the rotation procedure is proved for solving the complete problem of eigenvalues.

The exposition of the theory of bilinear and quadratic forms is concluded with the reduction of the equations of hypersurfaces of the second order in an n-dimensional space to a canonical form.

When tensors are considered, a tensor form of notation of the principal operations of vector algebra, important for applications, is presented in addition to the traditional material. The notions of pseudo-Euclidean space, Galilean coordinates and Lorentz' transformations are also presented.

The book is concluded with the exposition of the elements of the group theory and their representations.

It should be pointed out that this book is closely related to our *Analytic Geometry,* although it can be studied indepnntly.

The authors express their deep gratitude to A.N. Tikhonov and A.G. Sveshnikov for their valuable comments, to Sh.A. Alimov, whose contribution is far beyond ordinary editing, to L.D. Kudryavtsev, S.A. Lomov and especially to A.A. Samarsky for their valuable critical remarks and advice, to E.S. Nikolayev, D.D. Sokolov and E.V. Shikin for their help in writing some parts of the book.

V. Ilyin, E. Poznyak

INTRODUCTION

In this book we shall deal with seemingly different objects: (1) matrices (or rectangular tables of numbers), (2) algebraic forms including the so-called linear, bilinear and quadratic forms, (3) the so-called linear (Euclidean, in particular) spaces and linear transformations in those spaces.

The reader is sure to have a general idea of these objects. Indeed, square matrices of the second and third orders and the determinants corresponding to those matrices are studied in courses of analytic geometry. Linear and quadratic forms are, respectively, homogeneous linear and homogeneous quadratic functions of several independent variables (say, vector coordinates). We can cite the collection of all geometric vectors on a plane (or in space) with the defined operations of addition of those vectors and multiplication of them by numbers as an example of a linear space. If a scalar product is also given for the collection of the vectors, then we arrive at the notion of a Euclidean space. A transition from one rectangular Cartesian basis to another can serve as an example of a linear transformation in such a space.

Despite the seeming distinction, the collections of objects mentioned above are closely interrelated, most of the assertions admit of the same formulation for each of these collections. This interrelation becomes especially noticeable when we study arbitrary linear and Euclidean spaces (and linear transformations in those spaces).

However, a specified matrix approach to the results is directly connected with actual calculations (and, in particular, with the solution of linear systems of equations). That is why we begin our exposition with the consideration of matrices and often return to the matrix approach to the results obtained.

Chapter 1

MATRICES AND DETERMINANTS

In this chapter we consider tables of numbers which are called **matrices** and which play the most important role in what follows. The principal operations on matrices are introduced here and the properties of the so-called **determinants**, which are the main numerical characteristic of square matrices, are studied in detail.

1.1. Matrices

1.1.1. The concept of a matrix. A **matrix** is a rectangular array of numbers containing a certain number m of rows and a certain number n of columns.

The numbers m and n are the **orders** of a matrix. When $m=n$, the matrix is called **square** and the number $m=n$ is its **order.**

In what follows, a matrix will be written either between double lines or parentheses:

$$\begin{Vmatrix} a_{11} & a_{12} & & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ \cdot & \cdot & & \cdot \\ a_{m1} & a_{m2} & & a_{mn} \end{Vmatrix} \quad \text{or} \quad \begin{pmatrix} a_{11} & a_{12} & & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ & \cdot & & \cdot \\ a_{m1} & a_{m2} & & a_{mn} \end{pmatrix}.$$

For a brief designation of a matrix use will be often made either of a capital letter (say, A), or a symbol $\| a_{ij} \|$, sometimes followed by an elucidation: $A=\| a_{ij} \|=(a_{ij})\,(i=1, 2, \ldots, m;\ j=1, 2, \ldots, n)$.

The numbers a_{ij} constituting a matrix are called its **elements.** In the notation a_{ij} the first index i is the number of the row and the second index j is the number of the column.

In the case of a square matrix

$$\begin{Vmatrix} a_{11} & a_{12} & & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ & & & \\ a_{n1} & a_{n2} & & a_{nn} \end{Vmatrix} \tag{1.1}$$

the notions of a principal and a secondary diagonal are introduced. The **principal leading diagonal** of the matrix (1.1) is the diagonal $a_{11}\, a_{22} \ldots a_{nn}$ passing from the upper left corner of the matrix to its lower right corner. The **secondary diagonal** of the same matrix is the diagonal $a_{n1} a_{(n-1)2} \ldots a_{1n}$ passing from the lower left to the upper right corner.

1.1.2. The main operations on matrices and their properties. First of all, let us agree to consider two matrices to be **equal** if they have the same order and all their respective elements coincide.

We shall now define the main operations on matrices.

(a) **Addition of matrices.** The **sum** of two matrices $A = \| a_{ij} \|$ $(i=1, 2, \ldots, m;\ j=1, 2, \ldots, n)$ and $B = \| b_{ij} \|$ $(i=1, 2, \ldots, m;$ $j=1, 2, \ldots, n)$ of the same orders m and n is a matrix $C = \| c_{ij} \|$ $(i=1, 2, \ldots, m;\ j=1, 2, \ldots, n)$ of the same orders m and n, whose elements c_{ij} are

$$c_{ij} = a_{ij} + b_{ij}\ (i=1, 2, \ldots, m;\ j=1, 2, \ldots, n). \tag{1.2}$$

The sum of two matrices is designated as $C = A + B$. The operation of forming the sum of matrices is called **addition.**

Thus, by definition,

$$\begin{Vmatrix} a_{11} & a_{12} & & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ & \cdot & & \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{Vmatrix} + \begin{Vmatrix} b_{11} & b_{12} & & b_{1n} \\ b_{21} & b_{22} & & b_{2n} \\ \cdot & \cdot & & \\ b_{m1} & b_{m2} & & b_{mn} \end{Vmatrix}$$

$$= \begin{Vmatrix} (a_{11}+b_{11}) & (a_{12}+b_{12}) & & (a_{1n}+b_{1n}) \\ (a_{21}+b_{21}) & (a_{22}+b_{22}) & & (a_{2n}+b_{2n}) \\ \cdot & \cdot & \cdot & \cdot \\ (a_{m1}+b_{m1}) & (a_{m2}+b_{m2}) & \ldots & (a_{mn}+b_{mn}) \end{Vmatrix}.$$

It immediately follows from the definition of the sum of matrices, to be more precise, from formula (1.2), that the operation of matrix addition has the same properties as the operation of addition of real numbers, namely:

(1) commutativity: $A + B = B + A$.

(2) associativity: $(A + B) + C = A + (B + C)$.

These properties free us from being troubled about the sequence of matrices when we add two or more matrices.

(b) **Multiplication of a matrix by a number.** The **product of the matrix** $A = \| a_{ij} \|$ $(i=1, 2, \ldots, m;\ j=1, 2, \ldots, n)$ by a real number λ is a matrix $C = \| c_{ij} \|$ $(i=1, 2, \ldots, m;\ j=1, 2, \ldots, n)$ whose elements c_{ij} are

$$c_{ij} = \lambda a_{ij}\ (i=1, 2, \ldots, m;\ j=1, 2, \ldots, n). \tag{1.3}$$

The notation $C = \lambda A$ or $C = A\lambda$ is used for the product of a matrix

by a number. The operation of forming the product of a matrix by a number is called the **multiplication** of the matrix by that number.

It is clear from formula (1.3) that multiplication of a matrix by a number is subject to the following laws:

(1) associative law for the numerical factor:
$(\lambda\mu) A = \lambda (\mu A)$;

(2) distributive law for addition of matrices:
$\lambda (A+B) = \lambda A + \lambda B$;

(3) distributive law for the sum of the numbers: $(\lambda+\mu) A = \lambda A + \mu A$.

Remark. It is natural to term as the difference of two matrices A and B of the same orders m and n a matrix C of orders m and n which, being added to the matrix B, yields a matrix A. The natural notation for the difference of two matrices is $C = A - B$.

It is easy to verify that the difference C lwo matrices A and B can be obtained by the rule $C = A + (-1) \cdot B$.

(c) **Multiplication of matrices.** The **product** of the matrix $A = \| a_{ij} \|$ $(i=1, 2, \ldots, m;\ j=1, 2, \ldots, n)$, of the respective orders m and n, by the matrix $B = \| b_{ij} \|$ $(i=1, 2, \ldots, n;\ j=1, 2, \ldots, p)$ of the respective orders n and p is a matrix $C = \| c_{ij} \|$ $(i=1, 2, \ldots, m;\ j=1, 2, \ldots, p)$ of the respective orders m and p and ith elements c_{ij} defined by the formula

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} \quad (i=1, 2, \ldots, m; \quad j=1, \ldots, p). \tag{1.4}$$

The notation $C = A \cdot B$ is used to designate the product of the matrix A by the matrix B. The operation of forming the product of the matrix A by the matrix B is called the **multiplication** of those matrices.

It follows from the definition stated above that *the matrix A cannot be multiplied by any matrix B.* It is necessary that the number of columns of the matrix A be equal to the number of rows of the matrix B.

In particular, both products $A \cdot B$ and $B \cdot A$ can be defined only when the number of columns of A coincides with the number of rows of B and the number of rows of A coincides with the number of columns of B. In that case both matrices $A \cdot B$ and $B \cdot A$ are square but their orders, in a general case, are different. For the two products $A \cdot B$ and $B \cdot A$ to be not only defined but to have the same order, it is necessary and sufficient that both matrices A and B be square matrices of the same order.

Formula (1.4) is the rule of forming the elements of the matrix C which is the product of the matrix A by the matrix B. Here is the verbal formulation of this rule: *the element c_{ij} lying at the intersection of the ith row and the jth column ofematrix $C = A \cdot B$ is*

equal to the sum of the pairwise products of the respective elements of the ith row of the matrix A and the jth column of the matrix B.

We give the formula for multiplication of second-order square matrices as an example of application of the indicated rule:

$$\begin{Vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{Vmatrix} \cdot \begin{Vmatrix} b_{11} & b_{12} \\ a_{21} & b_{22} \end{Vmatrix} = \begin{Vmatrix} (a_{11}b_{11}+a_{12}b_{21}) & (a_{11}b_{12}+a_{12}b_{22}) \\ (a_{21}b_{11}+a_{22}b_{21}) & (a_{21}b_{12}+a_{22}b_{22}) \end{Vmatrix}.$$

Formula (1.4) yields the following properties of the product of the matrix A by the matrix B:

(1) associative law: $(AB)C=A(BC)$;

(2) distributive law for the sum of the matrices:
$(A+B)C=AC+BC$ or $A(B+C)=AB+AC$.

The distributive law follows immediately from formulas (1.4) and (1.2.), and to prove the associative law it is sufficient to note that if $A=\| a_{ij} \|$ $(i=1, 2, \ldots, m;\ j=1, 2, \ldots, n)$, $B=\| b_{jk} \|$ $(j=1, 2, \ldots, n;\ k=1, 2, \ldots, p)$, $C=\| c_{kl} \|$ $(k=1, 2, \ldots, p,\ l=1, 2, \ldots, r)$, then by virtue of (1.4) the element d_{il} of the matrix $(AB)C$ is equal to $d_{il}=\sum_{k=1}^{p}\left(\sum_{j=1}^{n} a_{ij}b_{jk}\right)\cdot c_{kl}$ and the element d'_{il} of the matrix $A(BC)$ is equal to $d'_{il}=\sum_{j=1}^{n} a_{ij}\left(\sum_{k=1}^{p} b_{jk}c_{kl}\right)$, but then the equality $d_{il}=d'_{il}$ follows from the possibility of changing the order of summation with respect to j and k.

It is reasonable to pose the question concerning the commutative law for the product of the matrix A by the matrix B only for square matrices A and B of the same order (since, as was mentioned above, only for such matrices A and B are the products AB and BA defined and are the matrices of the same order). Simple examples show that *the product of two square matrices of the same order does not possess, in general, a commutative law.* Indeed, if we set $A=\begin{Vmatrix} 0 & 1 \\ 0 & 0 \end{Vmatrix}$, $B=\begin{Vmatrix} 0 & 0 \\ 1 & 0 \end{Vmatrix}$ then $AB=\begin{Vmatrix} 1 & 0 \\ 0 & 0 \end{Vmatrix}$, and $BA=\begin{Vmatrix} 0 & 0 \\ 0 & 1 \end{Vmatrix}$.

There are, however, significant special cases in which the commutative law holds true*.

Among square matrices, there is a special class of **diagonal** ma-

* Two matrices whose product the commutative law holds true are customarily called **commuting matrices.**

trices in each of which the elements lying outside the principal diagonal are zero. Every diagonal matrix of order n has the form

$$D=\begin{Vmatrix} d_1 & 0 & & 0 \\ 0 & d_2 & & 0 \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \dots & d_n \end{Vmatrix}, \tag{1.5}$$

where $d_1, d_2, \dots, d_n$ are arbitrary numbers. It is easy to see that when all these numbers are equal, i.e. $d_1=d_2=\dots=d_n=d$, then the equality $AD=DA$ is valid for any square matrix A of order n. Indeed, let us designate as c_{ij} and c'_{ij} the elements lying at the intersection of the ith row and the jth column of the matrices AD and DA respectively. Then, from equation (1.4) and the form of the matrix D we get

$$c_{ij}=a_{ij}d_j=a_{ij}d, \quad c'_{ij}=d_ia_{ij}=da_{ij}, \tag{1.6}$$

i.e. $c_{ij}=c'_{ij}$.

Among all the diagonal matrices (1.5) with coinciding elements $d_1=d_2=\dots=d_n=d$, an especially important role is played by the following two matrices. The first matrix is obtained at $d=1$, is called an **identity** or **unit matrix** of order n and is denoted by I. The second matrix is obtained at $d=0$, is called a **null** or **zero matrix** of order n and is denoted by O. Thus,

$$I=\begin{Vmatrix} 1 & 0 & & 0 \\ 0 & 1 & & 0 \\ \cdot & \cdot & & \cdot \\ 0 & 0 & & 1 \end{Vmatrix}, \quad O=\begin{Vmatrix} 0 & 0 & & 0 \\ 0 & 0 & & 0 \\ \cdot & \cdot & & \cdot \\ 0 & 0 & & 0 \end{Vmatrix}.$$

By virtue of what was proved above, $AI=IA$ and $AO=OA$. Moreover, it is evident from (1.6) that

$$AI=IA=A, \quad AO=OA=O. \tag{1.7}$$

The first formula (1.7) characterizes the special role played by the identity matrix I, which is similar to that played by the number 1 in the multiplication of real numbers. As to the special role played by the null matrix O, it is revealed not only by the second formula (1.7) but also by the equality* $A+O=O+A=A$ which can be easily verified.

It should be pointed out in conclusion that the concept of a null matrix can be introduced for nonsquare matrices as well (a null matrix is *any* matrix whose all elements are zero).

1.1.3. Hypermatrices. Suppose that by horizontal and vertical lines a matrix $A=\|a_{ij}\|$ is partitioned into separate rectangular

* This equality is a direct consequence of formula (1.2).

cells each of which is a matrix of a smaller size and is called a **block** or a **submatrix** of the original matrix. In that case it is possible to regard the original matrix A as a certain new matrix $A = \| A_{\alpha\beta} \|$ (so-called **partitioned matrix** or **hypermatrix**) whose elements $A_{\alpha\beta}$ are the indicated blocks. Those elements are usually denoted by a capital letter to emphasize that they are, in general, matrices and not numbers and (like ordinary numerical elements) are supplied with two indices, the first indicating the number of the row of the block and the second indicating the number of the column.

For example, the matrix

$$A = \left\| \begin{array}{cccc|cc} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & a_{26} \\ \hline a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & a_{36} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} & a_{46} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} & a_{56} \end{array} \right\|$$

can be regarded as a hypermatrix $A = \left\| \begin{array}{cc} A_{11} & A_{12} \\ A_{21} & A_{22} \end{array} \right\|$, whose elements are the following blocks:

$$A_{11} = \left\| \begin{array}{cccc} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \end{array} \right\|, \quad A_{12} = \left\| \begin{array}{cc} a_{15} & a_{16} \\ a_{25} & a_{26} \end{array} \right\|,$$

$$A_{21} = \left\| \begin{array}{cccc} a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \\ a_{51} & a_{52} & a_{53} & a_{54} \end{array} \right\|, \quad A_{22} = \left\| \begin{array}{cc} a_{35} & a_{36} \\ a_{45} & a_{46} \\ a_{55} & a_{56} \end{array} \right\|.$$

It is remarkable that the principal operations on hypermatrices are performed according to the same rules that govern the operations on ordinary numerical matrices, with the only difference that blocks play the role of elements.

In fact, it can be easily verified that if the matrix $A = \| a_{ij} \|$ is a partitioned matrix with blocks $A_{\alpha\beta}$, then with the same partitioning, the matrix $\lambda A = \| \lambda a_{ij} \|$ is associated with the elements $\lambda A_{\alpha\beta}$*.

It can be as easily verified that if the matrices A and B are of the same order and are similarly partitioned, then the sum of the matrices A and B is associated with a hypermatrix with the elements $C_{\alpha\beta} = A_{\alpha\beta} + B_{\alpha\beta}$ (here $A_{\alpha\beta}$ and $B_{\alpha\beta}$ are the blocks of the matrices A and B).

Assume, finally, that A and B are two hypermatrices such that

* In that case the elements $\lambda A_{\alpha\beta}$ can be calculated by the rule of multiplication of the matrix $A_{\alpha\beta}$ by the number λ.

the number of columns in each block $A_{\alpha\beta}$ is equal to the number of rows of the block $B_{\beta\gamma}$ (so that the product $A_{\alpha\beta}B_{\beta\gamma}$ is defined for any α, β and γ). Then the product $C=AB$ is a matrix with the elements $C_{\alpha\gamma}$, defined by the formula

$$C_{\alpha\gamma}=\sum_{\beta} A_{\alpha\beta}B_{\beta\gamma}.$$

To prove this formula, it is sufficient to write out its left-hand and right-hand sides in the terms of ordinary (numerical) elements of the matrices A and B (we leave this to the reader).

As an example of the application of hypermatrices, we shall consider the notion of the so-called **direct sum** of square matrices.

The **direct sum** of two square matrices A and B of orders m and n respectively is a square hypermatrix C of order $m+n$ equal to $C=\begin{Vmatrix} A & O \\ O & B \end{Vmatrix}$. The notation $C=A\oplus B$ is used to denote the direct sum of the matrices A and B.

It is evident from the definition of the direct sum of the matrices A and B that this sum, in a general case, does not obey a commutative law. But the validity of its associative law can be easily verified: $(A\oplus B)\oplus C=A\oplus(B\oplus C)$.

The following formulas relating the operation of direct summation and the operations of ordinary addition and multiplication of matrices can be easily verified with the aid of the properties of operations on hypermatrices:

$$(A_m\oplus A_n)+(B_m\oplus B_n)=(A_m+B_m)\oplus(A_n+B_n),$$
$$(A_m\oplus A_n)(B_m\oplus B_n)=A_mB_m\oplus A_nB_n$$

(in these formulas A_m and B_m are arbitrary square matrices of order m, and A_n and B_n are arbitrary square matrices of order n). We invite the reader to verify these formulas.

1.2. Determinants

The aim of this section is to construct the theory of determinants of any order n. We assume that the reader is acquainted with the second- and third-order determinants and shall make no references in our exposition. The acquaintance with the second- and third-order determinants will make it easier for the reader to comprehend the material that follows.

1.2.1. The concept of a determinant. Let us consider an arbitrary

square matrix of any order n:

$$A = \begin{Vmatrix} a_{11} & a_{12} & & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ & & & \\ a_{n1} & a_{n2} & & a_{nn} \end{Vmatrix}. \tag{1.8}$$

We shall associate with every matrix of this kind a definite numerical characteristic, called a determinant, corresponding to that matrix.

If the order n of the matrix (1.8) is equal to unity, then the matrix consists of one element a_{11} and the value of that element is the first-order determinant corresponding to that matrix.

If the order n of the matrix (1.8) is equal to two, i.e. if the matrix has the form

$$A = \begin{Vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{Vmatrix}, \tag{1.9}$$

then the second-order determinant corresponding to that matrix is a number equal to $a_{11}a_{22} - a_{12}a_{21}$ and designated as* $\Delta = \det A = \begin{Vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{Vmatrix}$. Thus, by definition,

$$\Delta = \det A = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}. \tag{1.10}$$

Formula (1.10) is a rule of forming a second-order determinant from the elements of the matrix corresponding to it. The verbal formulation of this rule is the following: the second-order determinant corresponding to matrix (1.9) is equal to the difference between the product of the elements on the principal diagonal of that matrix and the product of the elements lying on its secondary diagonal**.

In what follows, we shall speak of the elements, rows and columns of a determinant meaning the elements, rows and columns, respectively, of the matrix corresponding to that determinant.

Let us now elucidate the concept of a determinant of **any order** n, where $n \geqslant 2$. We shall introduce the notion of such a determinant by induction, assuming that we have already introduced the concept of a determinant of order $n-1$, corresponding to an arbitrary square matrix of order $n-1$.

We shall agree to term as the ***minor of any element*** a_{ij} *of the nth-*

---

* As distinct from matrices, determinants are denoted by ordinary lines and not by double lines.

** Recall that the **principal** diagonal of a square matrix is the diagonal passing from the upper left corner to the lower right corner (i.e. $a_{11}a_{22}$ in the case of matrix (1.9)) and the **secondary** diagonal is that from the lower left to the upper right corner (i.e. $a_{21}a_{12}$ in the case of matrix (1.9)).

order matrix (1.8) *the determinant of order* $n-1$ *corresponding to the matrix obtained from matrix* (1.8) *by deleting the ith row and the jth column* (the row and the column at whose intersection the element a_{ij} is located). The minor of the element a_{ij} is designated as $\bar{M}_j^i$. The superscript denotes the number of the row, the subscript denotes the number of the column and the bar over the letter M denotes that the indicated row and column are deleted.

A **determinant of order** n *corresponding to matrix* (1.8) *is a number equal to* $\sum_{j=1}^{n}(-1)^{1+j}a_{ij}\bar{M}_j^1$ *and designated as*

$$\Delta = \det A = \begin{vmatrix} a_{11} & a_{12} & & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ \cdot & & & \\ a_{n1} & a_{n2} & & a_{nn} \end{vmatrix}. \tag{1.11}$$

Thus, by definition

$$\Delta = \det A = \begin{vmatrix} a_{11} & a_{12} & & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ \cdot & & & \cdot \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix} = \sum_{j=1}^{n}(-1)^{1+j}a_{1j}\bar{M}_j^1. \tag{1.12}$$

Formula (1.12) is the rule of forming a determinant of order n from the elements of the first row of the corresponding matrix and from the minors $\bar{M}_j^1$ of the elements of the first row which are determinants of order $n-1$.

Note that for $n=2$ rule (1.12) exactly coincides with rule (1.10) since in that case the minors of the elements of the first row have the form $\bar{M}_1^1 = a_{22}$, $\bar{M}_2^1 = a_{21}$.

The natural question arises whether it is possible to use the elements and the corresponding minors of an arbitrary ith row matrix (1.8) rather than of the first row in order to obtain the value of determinant (1.11). The following *fundamental* theorem provides an answer to that question.

Theorem 1.1. *Whatever the number of the row* i ($i=1, 2, \dots, n$), *the following formula* is valid for the nth-order determinant* (1.11):

$$\Delta = \det A = \sum_{j=1}^{n}(-1)^{i+j}a_{ij}\bar{M}_j^i, \tag{1.13}$$

called the expansion of that determinant according to the ith row.

Remark. It should be emphasized that in this formula the degree to which the number (-1) is raised is equal to the sum of the numbers

* By the meaning of the theorem, $n \geqslant 2$.

of the row and the column at whose intersection the element a_{ij} is located.

The proof of Theorem 1.1. Formula (1.13) should be proved only for the numbers $i=2, 3, \ldots, n$*. For $n=2$ (i.e. for a second-order determinant) this formula should be proved only for the number $i=2$, i.e. for $n=2$ it is only necessary to prove the formula

$$\Delta = \det A = \sum_{j=1}^{2} (-1)^{2+j} a_{2j} \bar{M}_j^2 = -a_{21}\bar{M}_1^2 + a_{22}\bar{M}_2^2.$$

The validity of the last formula follows immediately from the expressions for the minors of matrix (1.9) $\bar{M}_1^2 = a_{12}$, $\bar{M}_2^2 = a_{11}$, by virtue of which the right-hand side of this formula coincides with the right-hand side of (1.10). We have thus proved the theorem for $n=2$.

We shall carry out the proof of formula (1.13) for an arbitrary $n > 2$ by induction, i. e. we shall assume that the formula of the form (1.13) of the expansion according to any row is valid for a determinant of order $n-1$, and proceeding from this assumption we shall make sure that formula (1.13) is also valid for a determinant of order n.

To prove this, we shall need the notion of minors of matrix (1.8) of order $n-2$. The determinant of order $n-2$ corresponding to the matrix which can be obtained from matrix (1.8) by deleting two rows with the numbers i_1 and i_2 and two columns with the numbers j_1 and j_2 is called a **minor of order $n-2$** and is designated as $\bar{M}_{j_1 j_2}^{i_1 i_2}$.

An nth-order determinant Δ is introduced by formula (1.12), and in this formula every minor $\bar{M}_j^1$ is a determinant of order $n-1$ for which the formula of the form (1.13) of the expansion according to any row is valid by assumption.

Having fixed any number i ($i=2, 3, \ldots, n$), we expand every minor $\bar{M}_j^1$ in formula (1.12) *according to the ith row* of the principal determinant (1.11) (in the minor $\bar{M}_j^1$ itself that row is the $(i-1)$th).

As a result the whole determinant Δ will be represented as a certain linear combination** of $(n-2)$th-order minors $\bar{M}_{jk}^{1i}$ with the noncoinciding numbers j and k, i. e. in the form***

$$\Delta = \sum_{j=1}^{n} \sum_{k<j} \theta_{jk} \bar{M}_{jk}^{1i}. \tag{1.14}$$

To calculate θ_{jk} we note that the minor $\bar{M}_{jk}^{1i}$ results from the expan-

---

* Since for $i=1$ the right-hand side of (1.13) is equal to det A by definition.

** Recall that a linear combination of some quantities is the sum of the products of those quantities by certain real numbers.

*** Since the minor $\bar{M}_{jk}^{1i}$ coincides with $\bar{M}_{kj}^{1i}$ we shall take in turn all the $(n-2)$th-order minors with the given numbers of rows 1 and i varying j from 1 to n and taking all possible $k<j$ for every j.

sion according to the $(i-1)$ th row* *of only the following two minors of order $n-1$ that correspond to the elements of the first row of matrix* (1.8): the minor $\bar{M}_j^1$ and the minor $\bar{M}_k^1$ (since only these two minors of the elements of the first row contain **all the columns** of the minor $\bar{M}_{jk}^{1i}$).

In the expansion of the minors $\bar{M}_j^1$ and $\bar{M}_k^1$ according to the indicated $(i-1)$th row, we write out only those summands which contain the minor $\bar{M}_{jk}^{1i}$ (and denote by dots the other summands which are of no interest to us). Bearing in mind that the element a_{ik} of the minor $\bar{M}_j^1$ is at the intersection of the $(i-1)$th row and the $(k-1)$th column of that minor**, and the element a_{ij} of the minor $\bar{M}_k^1$ is at the intersection of the $(i-1)$th row and the jth column of that minor***, we get

$$\bar{M}_j^1=(-1)^{(i-1)+(k-1)}\, a_{ik}\bar{M}_{jk}^{1i}+\ldots, \tag{1.15}$$

$$\bar{M}_k^1=(-1)^{(i-1)+j}\, a_{ij}\bar{M}_{jkk}^{1i}+\ldots. \tag{1.16}$$

Substituting (1.15) and (1.16) into the right-hand side of (1.12) and collecting the coefficients in $\bar{M}_{jk}^{1i}$, we find that the multiplier θ_{jk} in (1.14) has the form

$$\theta_{jk}=(-1)^{(1+i+j+k)}[a_{1j}a_{ik}-a_{1k}a_{ij}]. \tag{1.17}$$

To complete the proof of the theorem, we shall show that the right-hand side of (1.13) is also equal to the sum appearing on the right-hand side of (1.14) with the same values of (1.17) for θ_{jk}.

For that purpose, we expand every minor $\bar{M}_j^i$ of order $n-1$ on the right-hand side of (1.13) **according to the first row.** As a result, the whole right-hand side of (1.13) will become a linear combination of the same minors $\bar{M}_{jk}^{1i}$

$$\sum_{j=1}^{n}\sum_{k<j}\theta_{jk}\bar{M}_{jk}^{1i} \tag{1.18}$$

with some coefficients θ_{jk}, and we have now to calculate the multipliers θ_{jk} and ascertain the validity of formula (1.17) for them.

To do that, we note that the minor $\bar{M}_{jk}^{1i}$ results from the expansion according to the first row of only the following two minors of order $n-1$ corresponding to the elements of the ith row of matrix (1.8): the minor $\bar{M}_j^i$ and the minor $\bar{M}_k^i$ (since only these two minors of the elements of the ith row contain **all the columns** of the minor $\bar{M}_{jk}^{1i}$).

---

* In matrix (1.8) that row will be the ith row.

** Since in the minor $\bar{M}_k^1$ the first row and the jth column of matrix (1.8) are absent and $j<k$.

*** Since in the minor $\bar{M}_k^1$ the first row of matrix (1.8) is absent and the only column of matrix (1.8) which is absent in this minos has the number $k>j$.

In the expansions of the minors $\bar{M}_j^i$ and $\bar{M}_k^i$ according to the first row, we write out only those summands which contain the minor $\bar{M}_{jk}^{1i}$ (and denote by dots the other summands which are of no interest to us). Taking into account that the element a_{1k} of the miror $\bar{M}_j^i$ lies at the intersection of the first row and the $(k-1)$th column* of that minor and the element a_{1j} of the minor $\bar{M}_k^i$ lies at the intersection of the first row and the jth column** of that minor, we get

$$\bar{M}_j^i=(-1)^{1+(k-1)}a_{1k}\cdot\bar{M}_{jk}^{1i}+\ldots, \tag{1.19}$$

$$\bar{M}_k^i=(-1)^{1+j}a_{1j}\cdot\bar{M}_{jk}^{1i}+\ldots. \tag{1.20}$$

Substituting (1.19) and (1.20) into the right-hand side of (1.13) and collecting the coefficients in $\bar{M}_{jk}^{1i}$, we find that θ_{jk} in the sum (1.18) is defined by the same formula (1.17) as in equation (1.14).

We have proved Theorem 1.1.

Theorem 1.1 has established the possibility of expanding an nth-order determinant according to any of its rows. A question now arises whether it is possible to expand a determinant of order n according to any of its **columns.** The following *fundamental* theorem provides a positive answer.

Theorem 1.2. *Whatever the number of the column j ($j=1,2,\ldots,n$), the following formula holds true for the nth-order determinant* (1.11):

$$\Delta=\det A=\sum_{i=1}^{n}(-1)^{i+j}a_{ij}\bar{M}_j^i. \tag{1.21}$$

This formula is known as the **expansion of the determinant according to the jth column.**

Proof. It is sufficient to prove the theorem for $j=1$, i.e. to derive an expansion formula according to the first column

$$\Delta=\sum_{i=1}^{n}(-1)^{i+1}a_{i1}\bar{M}_1^i. \tag{1.22}$$

When formula (1.22) is derived, then, to prove formulas (1.21) for any $j=2, 3, \ldots, n$, it is sufficient to interchange the roles of the rows and columns and to repeat word for word the arguments of Theorem 1.1.

Formula (1.22) can be derived by induction.

It is easy to verify this formula for $n=2$ (since for $n=2$

---

* Since $j<k$ and the jth column of matrix (1.8) is absent in the minor $\bar{M}_j^i$.

** Since $j<k$ and only the kth column of matrix (1.8) is absent in the minor $\bar{M}_k^i$.

the minors of the elements of the first column have the form $\bar{M}_1^1=a_{22}$, $\bar{M}_1^2=a_{12}$, it follows that for $n=2$ the right-hand side of (1.22) coincides with the right-hand side of (1.10)).

Let us assume that the formula for the expansion according to the first column (1.22) is valid for a determinant of order $n-1$ and, proceeding from this fact, ascertain the validity of that formula for a determinant of order n.

With this aim in view, we isolate the first term $a_{11}\bar{M}_1^1$ on the right-hand side of formula (1.12) for a determinant Δ of order n and expand the $(n-1)$th-order determinant $\bar{M}_j^1$ **according to the first column** in all the other terms.

As a result, formula (1.12) assumes the form

$$\Delta=a_{11}\bar{M}_1^1+\sum_{j=2}^{n}\sum_{i=2}^{n}\theta_{ij}\bar{M}_{1j}^{1i}, \tag{1.23}$$

where θ_{ij} are certain coefficients which have to be defined. To calculate θ_{ij}, we note that the minor $\bar{M}_{1j}^{1i}$ results from the expansion according to the first column of only one of the minors of order $n-1$ corresponding to the first row*, of the minor $\bar{M}_j^1$. In the expansion of the minor $\bar{M}_i^1$ (for $j\geqslant 2$) according to the first column we write only that term which contains the minor $\bar{M}_{1j}^{1i}$ (and use dots to denote the other terms which are of no interest to us). Taking into account that the element a_{i1} of the minor $\bar{M}_j^1$ (for $j\geqslant 2$) lies in the $(i-1)$th row and the first column of that minor, we find that for $j\geqslant 2$

$$\bar{M}_j^1=(-1)^{(i-1)+1}a_{i1}\bar{M}_{1j}^{1i}+\ldots \tag{1.24}$$

Substituting (1.24) into the right-hand side of (1.12) (from which the first term has been deleted) and collecting the coefficients in $\bar{M}_{1j}^{1i}$, we find that the coefficient θ_{ij} in (1.23) has the form

$$\theta_{ij}=(-1)^{i+j+1}a_{1j}a_{i1}. \tag{1.25}$$

It remains to prove that the right-hand side of (1.22) is also equal to the sum appearing on the right-hand side of (1.23) with the same values (1.25) for θ_{ij}.

For that purpose, we isolate the first term $a_{11}\bar{M}_1^1$ on the right-hand side of (1.22) and expand the $(n-1)$th-order minor $\bar{M}_1^i$ **according to the first row** in all other terms.

As a result, the right-hand side of (1.22) will be the sum of the first term $a_{11}\bar{M}_1^1$ and a linear combination with some coefficients

* The minor $\bar{M}_1^1$ is assumed to be excluded.

θ_{ij} of the minors $\bar{M}_{1j}^{1i}$ of order $n-2$, i.e. will have the form

$$a_{11}\bar{M}_1^1+\sum_{i=2}\sum_{j=2}^{n}\theta_{ij}\bar{M}_{1j}^{1i}, \tag{1.26}$$

and we have only to calculate the terms θ_{ij} and ascertain the validity of formula 1.25) for them.

To do that, we note that the minor $\bar{M}_{1j}^{1i}$ results from the expansion, according to the first row, of only one of the $(n-1)$th-order minors corresponding to the first column, the minor $\bar{M}_1^i$. In the expansion of the minor $\bar{M}_1^i$(for $i\geqslant 2$) according to the first row we write only that term which contains the minor $\bar{M}_{1j}^{1i}$ (and denote by dots the other terms which are of no interest to us). Taking into account that the element a_{1j} of the minor $\bar{M}_1^i$ is in the first row and the $(j-1)$th column of that minor, we find that for $i\geqslant 2$

$$\bar{M}_1^i=(-1)^{1+(j-1)}a_{1j}\bar{M}_{1j}^{1i}+\ldots. \tag{1.27}$$

Substituting (1.27) *into the right-hand side of* (1.22) *from which the first term has been deleted, and forming a coefficient in* $\bar{M}_{1j}^{1i}$, by collecting the terms we find that θ_{ij} in the sum (1.26) is defined by the same formula (1.25) as in equation (1.23). We have proved Theorem 1.2.

1.2.2. Expressing a determinant directly in terms of its elements. Let us derive a formula making it possible to express a determinant of order n directly in terms of its elements (by-passing the minors).

Suppose each number $\alpha_1, \alpha_2, \ldots, \alpha_n$ assumes one of the values $1, 2, \ldots, n$, none of these numbers coinciding (in that case we say that the numbers $\alpha_1, \alpha_2, \ldots, \alpha_n$ are a certain **transposition** of the numbers $1, 2, \ldots, n$). We form all possible pairs $\alpha_i\alpha_j$ of the numbers $\alpha_1, \alpha_2, \ldots, \alpha_n$ and say that the pair $\alpha_i\alpha_j$ forms a **disorder** if $\alpha_i>\alpha_j$ for $i<j$. The total number of disorders formed by these pairs which can be compiled from the numbers $\alpha_1, \alpha_2, \ldots, \alpha_n$ will be symbolized as $N(\alpha_1, \alpha_2, \ldots, \alpha_n)$.

Next we derive by induction the following formula for the nth-order determinant (1):

$$\Delta=det\,A=\sum_{\alpha_1,\alpha_2,\ldots,\alpha_n}(-1)^{N(\alpha_1,\alpha_2,\ldots,\alpha_n)}a_{\alpha_1 1}a_{\alpha_2 2}\ldots a_{\alpha_n n} \tag{1.28}$$

(in this formula the summation extends over all possible interchanges (transpositions) $\alpha_1, \alpha_2, \ldots, \alpha_n$ of the numbers $1, 2, \ldots, n$; the number of these transpositions is evidently $n!$).

Formula (1.28) can be easily verified when $n=2$ (then only two transpositions, 1, 2 and 2, 1 are possible and, since $N(1, 2)=0$, $N(2, 1)=1$, formula (1.28) turns into equation (1.10)).

To prove the formula by induction, we assume that at $n>2$ formula (1.28) holds true for a determinant of order $n-1$.

Then, writing the expansion of the nth-order determinant (1.11) according to the first column*

$$\Delta=\det A=\sum_{\alpha_1=1}^{n}(-1)^{\alpha_1+1}a_{\alpha_1 1}\bar{M}_1^{\alpha_1}, \tag{1.29}$$

we can, by virtue of the induction hypothesis, represent each minor $\bar{M}_1^{\alpha_1}$ order $n-1$ in the form

$$\bar{M}_1^{\alpha_1}=\sum_{\alpha_2,\ldots,\alpha_n}(-1)^{N(\alpha_2,\ldots,\alpha_n)}a_{\alpha_2 2}\ldots a_{\alpha_n n} \tag{1.30}$$

(the summation extends over all possible transpositions $\alpha_2, \ldots, \alpha_n$ of $(n-1)$ natural numbers from 1 to n with the exception of the number α_1).

Since besides the pairs formed from the numbers $\alpha_2, \ldots, \alpha_n$ we can form only the pairs $\alpha_1\alpha_2, \alpha_1\alpha_3, \ldots, \alpha_1\alpha_n$ from the numbers $\alpha_1, \alpha_2, \ldots, \alpha_n$ and since among the numbers $\alpha_2, \ldots, \alpha_n$ there are exactly (α_1-1) numbers smaller than the number α_1, it follows that $N(\alpha_1, \alpha_2, \ldots, \alpha_n)=N(\alpha_2, \ldots, \alpha_n)+\alpha_1-1$.

Hence $(-1)^{N(\alpha_2,\ldots,\alpha_n)}(-1)^{\alpha_1+1}=(-1)^{N(\alpha_1,\alpha_2,\ldots,\alpha_n)}$ and, substituting (1.30) into (1.29), we exactly obtain formula (1.28). We have thus completed the derivation of formula (1.28).

It should be pointed out in conclusion that in the majority of courses of linear algebra formula (1.28) underlies the notion of an nth-order determinant.

1.2.3. Laplace's expansion.** We shall derive here a remarkable formula generalizing the formula for the expansion of a determinant of order n according to some row.

For that purpose we introduce into consideration **two types** of minors of a matrix of order n (1.8).

Suppose k is any number smaller than n and $i_1, i_2, \ldots, i_k$ and $j_1, j_2, \ldots, j_k$ are arbitrary numbers satisfying the conditions $1\leqslant i_1<i_2<\ldots<i_k\leqslant n$, $1\leqslant j_1<j_2<\ldots<j_k\leqslant n$.

Minors of the **first** type $M_{j_1 j_2\ldots j_k}^{i_1 i_2\ldots i_k}$ are determinants of order k corresponding to the matrix formed by the elements of matrix (1.8) lying at the intersection of k rows labelled $i_1, i_2, \ldots, i_k$ and k columns labelled $j_1, j_2, \ldots, j_k$.

Minors of the **second** type $\bar{M}_{j_1 j_2\ldots j_k}^{i_1 i_2\ldots i_k}$ are determinants of order $n-k$ corresponding to the matrix obtained from matrix (1.8) by

* This time it is convenient for us to designate the index of summation as α_1.

** Laplace, Pierre Simon, (1749-1827), an eminent French astronomer, mathematician and physicist.

deleting k rows labelled $i_1, i_2, \ldots, i_k$ and k columns labelled $j_1, j_2, \ldots, j_k$.

Minors of the second type are naturally called **complementary** with respect to minors of the first type.

Theorem 1.3 (Laplace's expansion). *For any number k smaller than n and for any fixed numbers of rows $i_1, i_2, \ldots, i_k$ such that $1 \leqslant i_1 < i_2 < \ldots < i_k \leqslant n$, the following formula holds true for the determinant of order n* (1.11):

$$\Delta = \det A = \sum_{j_1, j_2, \ldots, j_k} (-1)^{i_1 + \cdots + i_k + j_1 + \cdots + j_k} \times M_{j_1 j_2 \ldots j_k}^{i_1 i_2 \ldots i_k} \bar{M}_{j_1 j_2 \ldots j_k}^{i_1 i_2 \ldots i_k}. \tag{1.31}$$

This formula is called the **expansion of the determinant according to k rows** $i_1, i_2, \ldots, i_k$. *The summation in this formula is carried out by all possible values of the indices $j_1, j_2, \ldots, j_k$ satisfying the conditions $1 \leqslant j_1 < j_2 < \ldots j_k \leqslant n$.*

Proof. Note first of all that formula (1.31) is a generalization of the formula we have proved for the expansion of a determinant of order n according to **one** of its rows labelled i_1 into which it passes at $k=1$ (in that case the minor $M_{j_1}^{i_1}$ coincides with the element $a_{i_1 j_1}$ and minor $\overline{M}_{j_1}^{i_1}$ is the minor of the element $a_{i_1 j_1}$ introduced above).

Thus we have proved formula (1.31) for $k=1$. We shall use induction to prove this formula for any k satisfying the inequalities $1 < k < n$, i.e. we shall assume that formula (1.31) is valid for $k-1$ rows and proceeding from this fact ascertain the validity of formula (1.31) for k rows.

Thus suppose that $1 < k < n$ and that any k rows of matrix (1.8) labelled $i_1, i_2, \ldots, i_k$ and satisfying the conditions $1 \leqslant i_1 < i_2 < \ldots < i_k \leqslant n$ are fixed. Then, by assumption, the following formula holds true for $k-1$ rows labelled $i_1, \ldots, i_{k-1}$:

$$\Delta = \sum_{j_1, \ldots, j_{k-1}} (-1)^{i_1 + \ldots + i_{k-1} + j_1 + \cdots + j_{k-1}} \times M_{j_1 \ldots j_{k-1}}^{i_1 \ldots i_{k-1}} \bar{M}_{j_1 \ldots j_{k-1}}^{i_1 \ldots i_{k-1}} \tag{1.32}$$

(the summation here is taken over all possible values of the indices $j_1, \ldots, j_{k-1}$ satisfying the conditions $1 \leqslant j_1 < j_2 < \ldots < j_{k-1} \leqslant n$).

Let us expand every minor $\bar{M}_{j_1 \ldots j_{k-1}}^{i_1 \ldots i_{k-1}}$ in formula (1.32) according to the row labelled i_k in matrix (1.8). As a result, the whole determinant Δ will be represented as some linear combination of the minors $\bar{M}_{j_1 \ldots j_{k-1} j_k}^{i_1 \ldots i_{k-1} i_k}$ with coefficients which we shall designate

as $\theta_{j_1 \ldots j_k}$, i.e. the equality* $\Delta = \sum_{j_1, \ldots, j_k} \theta_{j_1 \ldots j_k} \bar{M}^{i_1 \ldots i_k}_{j_1 \ldots j_k}$ will hold true for Δ and it remains to calculate the coefficients $\theta_{j_1 \ldots j_k}$ and ascertain that they are equal to

$$\theta_{j_1 \ldots j_k} = (-1)^{i_1 + \ldots + i_k + j_1 + \ldots + j_k} M^{i_1 \ldots i_k}_{j_1 \ldots j_k}. \tag{1.33}$$

With that aim in view, we note that the $(n-k)$th minor $\bar{M}^{i_1 \ldots i_k}_{j_1 \ldots j_k}$ results from the expansion, according to the row labelled i_k, of only the following k minors of the $(n-k+1)$th order:

$$\bar{M}^{i_1 \ldots i_{k-1}}_{j_1 \ldots j_k \text{ (without } j_s)}{}^{**} \quad (s = 1, 2, \ldots, k), \tag{1.34}$$

since none of the other minors of order $n-k+1$ containing the row i_s, contains all the rows and all the columns of the minor $\bar{M}^{i_1 \ldots i_k}_{j_1 \ldots j_k}$.

In the expansion of every minor (1.34) according to the row labelled i_k of matrix (1.8), we write out only the term containing the minor $\bar{M}^{i_1 \ldots i_k}_{j_1 \ldots j_k}$ (and denote by dots all the other terms which are of no interest to us). Taking into account that in every minor (1.34) the element $a_{i_k j_s}$ lies in the row $[i_k - (k-1)]$ and the column $[j_s - (s-1)]$ of that minor**, we get

$$\bar{M}^{i_1 \ldots i_{k-1}}_{j_1 \ldots j_k \text{ (without } j_s)} = (-1)^{[i_k - (k-1)] + [j_s - (s-1)]} a_{i_k j_s} \bar{M}^{i_1 \ldots i_k}_{j_1 \ldots j_k} + \ldots .$$

We have now only to bear in mind that every minor (1.34) in formula (1.32) must be multiplied by the factor

$$(-1)^{(i_1 + \ldots + i_{k-1}) + (j_1 + \ldots + j_k) - j_s} M^{i_1 \ldots i_{k-1}}_{j_1 \ldots j_k \text{ (without } j_s)}$$

and then the sum is taken over all s from 1 to k. Taking also

---

* As before, the summation in this equation is taken over all possible values of the indices $j_1, \ldots, j_k$ satisfying the condition $1 \leqslant j_1 < j_2 < \ldots \leqslant j_k \leqslant n$.

** The designation $M^{i_1 \ldots i_{k-1}}_{j_1 \ldots j_k \text{(without } j_s)}$ is used for the minor corresponding to the intersection of the rows labelled $i_1, \ldots, i_{k-1}$ and the columns labelled $j_1, j_2, \ldots, j_k$, except for the column labelled j_s, and the designation (1.34) is used for the complementary minor.

*** It follows from the fact that the row labelled i_k is preceded by $k-1$ rows, and the column labelled j_s is preceded by $s-1$ columns of the minor $M^{i_1 \ldots i_{k-1}}_{j_1 \ldots j_k \text{(without } j_s)}$ of which the minor (1.34) is complementary.

into consideration that $(-1)^{2-2k-2s}=1$, we find that

$$\theta_{j_1\ldots j_k}=(-1)^{i_1+\ldots+i_k+j_1+\ldots+j_k}\left[\sum_{s=1}^{k}(-1)^{k+s}M^{i_1\ldots i_{k-1}}_{j_1\ldots j_k(\text{without } j_s)}\right].$$

Noting that the sum in brackets is the expansion of the minor $M^{i_1\ldots i_k}_{j_1\ldots j_k}$ according to its last, kth, row, we finally get formula (1.33) for $\theta_{j_1\ldots j_k}$. We have completed the proof of Laplace's theorem.

Remark. The formula for the expansion of a determinant according to some k columns is written and derived by a complete analogy with formula (1.32).

1.2.4. Properties of determinants. Given below are some properties possessed by an arbitrary nth-order determinant.

1°. The property of the equivalence of rows and columns. The **transposition** of any matrix or determinant is an operation as a result of which the rows and columns are interchanged but their sequence is retained. The transposition of a matrix A results in a matrix **transposed with respect to the matrix** A and designated as A'.

In what follows we agree to use the symbols $|A|$, $|B|$, $|A'|$ ... for the determinants of the square matrices A, B, A' ... respectively.

The first property of a determinant is formulated as follows: *in a transposition the value of a determinant is retained, i.e.* $|A'|=|A|$.

This property follows directly from Theorem 1.2 (it is sufficient to note that the expansion of the determinant $|A|$ according to the first column identically coincides with that of the determinant $|A'|$ according to the first row).

The property we have proved signifies a complete equivalence of the rows and columns and enables us to establish all the other properties **only for the rows** and to be sure of their validity for the columns.

2°. The property of antisymmetry in interchanging two rows (or two columns). *When two rows (or two columns) are interchanged, the determinant retains its absolute value but changes its sign to the opposite.*

For a second-order determinant, this property can be easily verified (it follows immediately from (1.10) *that the determinants* $\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$ and $\begin{vmatrix} a_{21} & a_{22} \\ a_{11} & a_{12} \end{vmatrix}$ differ only in sign).

Assuming that $n>2$, we shall consider the nth-order determinant (1.11) and suppose that two rows labelled i_1 and i_2 are interchanged in that determinant. Writing Laplace's formula for expansion

according to those two rows, we have

$$\Delta = \sum_{j_1, j_2} (-1)^{i_1+i_2+j_1+j_2} M_{j_1 j_2}^{i_1 i_2} \bar{M}_{j_1 j_2}^{i_1 i_2}. \tag{1.35}$$

When the rows labelled i_1 and i_2 are interchanged, every second-order determinant $M_{j_1 j_2}^{i_1 i_2}$ changes sign to the opposite by virtue of what was proved above, and all the other quantities appearing under the summation sign in (1.35) do not depend on the elements of the rows labelled i_1 and i_2 and retain their values. We have thus proved property 2°.

3°. **Linear property of a determinant.** *We say that some row* $(a_1, a_2, \ldots, a_n)$ *is a* ***linear combination*** *of the rows* $(b_1, b_2, \ldots, b_n)$ $(c_1, c_2, \ldots, c_n), \ldots, (d_1, d_2, \ldots, d_n)$ *with coefficients* $\lambda, \mu, \ldots, \nu$ *if* $a_j = \lambda b_j + \mu c_j + \ldots + \nu d_j$ *for all* $j = 1, 2, \ldots, n$.

The linear property of a determinant can be formulated as follows: *if in the nth-order determinant* Δ *a certain ith row* $(a_{i1}, a_{i2}, \ldots, a_{in})$ *is a linear combination of two rows* $(b_1, b_2, \ldots, b_n)$ *and* $(c_1, c_2, \ldots, c_n)$ *with coefficients* λ *and* μ, *then* $\Delta = \lambda\Delta_1 + \mu\Delta_2$, *where* Δ_1 *is a determinant whose ith row is equal to* $(b_1, b_2, \ldots, b_n)$ *and all the other rows are the same as in* Δ, *and* Δ_2 *is a determinant whose ith row is equal to* $(c_1, c_2, \ldots, c_n)$ *and all the other rows are the same as in* Δ.

To prove this property, let us expand each of the three determinants Δ, Δ_1 and Δ_2 according to the ith row and note that in all the three determinants all the minors $\bar{M}_j^i$ of the elements of the ith row are the same. Hence, the formula $\Delta = \lambda\Delta_1 + \mu\Delta_2$ immediately follows from the equations $a_{ij} = \lambda b_j + \mu c_j$ $(j = 1, 2, \ldots, n)$.

The linear property is sure to be valid also for the case when the ith row is a linear combination not of two but of several rows. The linear property is also valid for the columns of the determinant.

The three properties we have proved are the **principal** properties of a determinant revealing its nature.

The following five properties are **logical consequences** of the three principal properties.

Corollary 1. *A determinant with two similar rows (or columns) is equal to zero.* (Indeed, when two similar rows are interchanged, the determinant Δ does not change, on the one hand, and, on the other hand, it changes sign to the opposite by virtue of property 2°. Thus $\Delta = -\Delta$, i.e. $2\Delta = 0$ or $\Delta = 0$.)

Corollary 2. *Multiplication of all elements of a certain row (or a certain column) of a determinant by the number* λ *is equivalent to the multiplication of the determinant by that number* λ.

In other words, *the common multiplier of all the elements of a certain row (or a certain column) of a determinant can be*

put before the sign of that determinant. (This property follows from property 3° for $\mu=0$.)

Corollary 3. *If all the elements of some row (or some column) of a determinant are zero, the determinant itself is equal to zero.* (Thiis property follows from property 2° for $\lambda=0$.)

Corollary 4. *If the elements of two rows (or two columns) of a determinant are proportional, the determinant is equal to zero.* (Indeed, by virtue of Corollary 2 the proportionality factor can be put before the sign of the determinant, and then a determinant with two identical rows remains which is equal to zero in accordance with Corollary 1.)

Corollary 5. *If we add to the elements of some row (or some column) of a determinant the respective elements of another row (another column) multiplied by an arbitrary factor λ, the value of the determinant does not change.* (In fact, the determinant resulting from the indicated addition can be represented, by virtue of property 3°, as the sum of two determinants the first of which coincides with the original determinant and the second is equal to zero because of the proportionality of the two rows (or columns) and Corollary 4.)

Remark. Corollary 5, as well as the linear property, admits of a more general formulation which we give for the rows: *if we add to the elements of some row of a determinant the respective elements of the row which is a linear combination of several other rows of that determinant (with arbitrary coefficients) the value of the determinant does not change.*

Corollary 5 is widely used in actual calculations of determinants (the corresponding examples are given in 1.2.5).

Before formulating one more property of determinants, we introduce a useful notion of an algebraic adjunct (or a cofactor or a signed minor) of the given element of a determinant.

*An **algebraic adjunct** of the given element a_{ij} of the nth-order determinant* (1.11) *is a number equal to* $(-1)^{i+j}\bar{M}_j^i$ *and designated as* A_{ij}.

Thus, the algebraic adjunct of a given element may differ from the minor of that element only in sign.

Using the notion of an algebraic adjunct, we can reformulate theorems 1.1 aand 1.2 as follows: *the sum of the products of the elements of any row (any column) of a determinant by the corresponding algebraic adjuncts of that row (column) is equal to that determinant.*

The corresponding formulas for the expansion of a determinant according to the ith row and the jth column can be rewritten

as follows:

$$\Delta = \sum_{j=1}^{n} a_{ij}A_{ij} \text{ (for any } i=1, 2, \ldots, n), \tag{1.13'}$$

$$\Delta = \sum_{i=1}^{n} a_{ij}A_{ij} \text{ (for any } j=1, 2, \ldots, n). \tag{1.21'}$$

We can now formulate the last property of a determinant.

4°. The property of algebraic adjuncts of adjacent rows (or columns). *The sum of the products of the elements of some row (or some column) of a determinant by the corresponding algebraic adjuncts of the elements of any other row (any other column) is equal to zero.*

We shall carry out the proof for the rows (it is quite analogous for the columns). Writing out formula (1.13′)

$$\begin{vmatrix} a_{11} & a_{12} & & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ \cdot & & & \cdot \\ a_{n1} & a_{n2} & & a_{nn} \end{vmatrix} = A_{i1}a_{i1} + A_{i2}a_{i2} + \ldots + A_{in}a_{in} \tag{1.36}$$

we note that since the algebraic adjuncts $A_{i1}, A_{i2}, \ldots, A_{in}$ are independent of the elements of the ith row $a_{i1}, a_{i2}, \ldots, a_{in}$, equation (1.36) is an identity with respect to $a_{i1}, a_{i2}, \ldots, a_{in}$ and is retained when the numbers $a_{i1}, a_{i2}, \ldots, a_{in}$ are replaced by any other n numbers. Replacing $a_{i1}, a_{i2}, \ldots, a_{in}$ by the respective elements of any (different from the ith) kth row $a_{k1}, a_{k2}, \ldots, a_{kn}$, we get, on the left of (1.36), a determinant with two similar rows which is equal to zero in accordance with property 1. Thus,

$$A_{i1}a_{k1} + A_{i2}a_{k2} + \ldots + A_{in}a_{kn} = 0$$

(for any noncoinciding i and k).

1.2.5. Examples of calculations of determinants. When calculating determinants, we make wide use of formulas for expansion according to a row and a column and Corollary 5 which enables us, without changing the value of the determinant, to add to any row (or column) an arbitrary linear combination of its other rows (or columns). It is especially convenient to use the formula for expansion according to the rows (or columns) many of whose elements are zero. In particular, if only one element of the given row is different from zero, then the expansion according to that row contains only one term and reduces the problem of calculating a determinant of order n to that of calculating a determinant of order $n-1$ (of the minor contained in the indicated term).

If, in the given row, several elements corresponding to the

intersection of that row with several columns are different from zero, then, applying Corollary 5 to the indicated columns, we can turn into zero all the elements of that row, except one without changing the determinant.

Let us consider some examples.

Example 1. Suppose we have to calculate the following determinant of order 4:

$$\Delta = \begin{vmatrix} 4 & 99 & 83 & 1 \\ 0 & 8 & 16 & 0 \\ 60 & 17 & 134 & 20 \\ 15 & 43 & 106 & 5 \end{vmatrix}.$$

Subtracting the trebled last column from the first column, we get

$$\Delta = \begin{vmatrix} 1 & 99 & 83 & 1 \\ 0 & 8 & 16 & 0 \\ 0 & 17 & 134 & 20 \\ 0 & 43 & 106 & 5 \end{vmatrix}.$$

It is natural then to expand the determinant according to the first column. As a result we get

$$\Delta = 1 \cdot \begin{vmatrix} 8 & 16 & 0 \\ 17 & 134 & 20 \\ 43 & 106 & 5 \end{vmatrix}.$$

Now, in the third-order determinant, we subtract the doubled first column from the second column and get

$$\Delta = \begin{vmatrix} 8 & 0 & 0 \\ 17 & 100 & 20 \\ 43 & 20 & 5 \end{vmatrix}.$$

Expanding now the last determinant of order 3 according to the first row, we finally get

$$\Delta = 8 \cdot \begin{vmatrix} 100 & 20 \\ 20 & 50 \end{vmatrix} = 8(500-400) = 800.$$

Example 2. Let us calculate the so-called *triangular determinant* whose all elements lying above the principal diagonal are zero:

$$\Delta_n = \begin{vmatrix} a_{11} & 0 & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 \\ \cdot & & & \cdot \\ a_{(n-1)1} & a_{(n-1)2} & a_{(n-1)(n-1)} & 0 \\ a_{n1} & a_{n2} & a_{n(n-1)} & a_{nn} \end{vmatrix}.$$

Expanding the determinant Δ_n according to the last column, we find that it is equal to the product of the element a_{nn} by the $(n-1)$th-order triangular determinant Δ_{n-1} equal to

$$\Delta_{n-1}=\begin{vmatrix} a_{11} & 0 & & 0 \\ a_{21} & a_{22} & & 0 \\ \cdot & \cdot & \cdot\ \cdot\ \cdot & \cdot \\ a_{(n-1)1} & a_{(n-1)2} & \cdots & a_{(n-1)(n-1)} \end{vmatrix}.$$

We again expand the last determinant according to the last column and make sure that it is equal to the product of the element $a_{(n-1)(n-1)}$ by the $(n-2)$th-order triangular determinant Δ_{n-2}. Continuing similar arguments we arrive at the following expression for the original determinant: $\Delta_n=a_{11}a_{22}\ldots a_{nn}$.

Thus, *a triangular determinant is equal to the product of the elements lying on its principal diagonal.*

Remark 1. If all the elements of the determinant Δ lying below its principal diagonal are equal to zero, then the determinant is also equal to the product of the elements lying on its principal diagonal (we can ascertain this using the scheme presented above, applying it, however, not to the last columns but to the last rows, or we can simply transpose Δ and reduce this case to that considered above).

We can find, in a similar way, that a determinant whose all elements lying above (or below) the secondary diagonal are zero is equal to the product of the number $(-1)^{n(n-1)/2}$ by all the elements lying on that diagonal.

Example 3. The determinant of order $2n$ of the hypermatrix $\begin{Vmatrix} A & O \\ B & C \end{Vmatrix}$ in which A, B and C are arbitrary square matrices of order n and O is a null square matrix of order n, can serve as a generalization of a second-order triangular determinant. Let us ascertain that the following formula is valid for the indicated determinant:

$$\begin{vmatrix} A & O \\ B & C \end{vmatrix} = |A|\ |C|^{*}. \tag{1.37}$$

Applying Laplace's expansion, let us expand the determinant appearing on the left-hand side of (1.37) *according to the first n rows.* Since a determinant in which at least one column consists of zeros is equal to zero it follows, according to (1.31), that only one term is different from zero, and that term (by virtue

* Recall that we have agreed to use the designations $|A|$, $|B|$, $|C|$ for the determinants of the matrices A, B, C ... respectively.

of the fact that $(-1)^{(1+\ldots+n)+(1+\ldots+n)}=1$) is precisely equal to $|A|\,|C|$.

Remark 2. Using similar arguments, we can verify the validity of the formula

$$\begin{vmatrix} A & B \\ C & O \end{vmatrix} = (-1)^n\,|B|\,|C| \tag{1.38}$$

(A, B, C and O have the same sense as before).

For that purpose we must expand the determinant appearing on the left-hand side of (1.38) *according to the last n rows* and take into account that

$$(-1)^{[(n+1)+\ldots+2n]+[1+\ldots+n]}=(-1)^{2n(2n+1)/2}=(-1)^n.$$

Example 4. Let us now calculate the so-called **Vandermonde determinant**

$$\Delta(x_1, x_2, \ldots, x_n)=\begin{vmatrix} 1 & 1 & & 1 \\ x_1 & x_2 & & x_n \\ x_1^2 & x_2^2 & & x_n^2 \\ \cdot & \cdot & & \cdot \\ x_1^{n-1} & x_2^{n-1} & \ldots & x_n^{n-1} \end{vmatrix}. \tag{1.39}$$

Subtracting the first column from all the rest columns, we get

$$\Delta(x_1, x_2, \ldots, x_n)=\begin{vmatrix} 1 & 0 & \ldots & 0 \\ x_1 & (x_2-x_1) & & (x_n-x_1) \\ x_1^2 & (x_2^2-x_1^2) & & (x_n^2-x_1^2) \\ \cdot & \cdot & & \cdot \\ x_1^{n-1} & (x_2^{n-1}-x_1^{n-1}) & & (x_n^{n-1}-x_1^{n-1}) \end{vmatrix}.$$

It is natural now to expand the determinant according to the first row. As a result we get

$$\Delta(x_1, x_2, \ldots, x_n)=\begin{vmatrix} (x_2-x_1) & (x_3-x_1) & \ldots & (x_n-x_1) \\ (x_2^2-x_1^2) & (x_3^2-x_1^2) & \ldots & (x_n^2-x_1^2) \\ \cdot & \cdot & & \cdot \\ (x_2^{n-1}-x_1^{n-1}) & (x_3^{n-1}-x_1^{n-1}) & & (x_n^{n-1}-x_1^{n-1}) \end{vmatrix}.$$

Subtracting now from each row the preceding row multiplied by x_1, we obtain

$$\Delta(x_1, x_2, \ldots, x_n)=\begin{vmatrix} (x_2-x_1) & (x_3-x_1) & (x_n-x_1) \\ x_2(x_2-x_1) & x_3(x_3-x_1) & x_n(x_n-x_1) \\ \cdot & \cdot & \cdot \\ x_2^{n-2}(x_2-x_1) & x_3^{n-2}(x_3-x_1) & x_n^{n-2}(x_n-x_1) \end{vmatrix}.$$

Next we can put before the sign of the determinant the common

factor of the first column equal to (x_2-x_1), the common factor of the second column equal to (x_3-x_1), ..., the common factor of the $(n-1)$th column equal to (x_n-x_1). As a result we obtain

$$\Delta(x_1, x_2, \ldots, x_n)=(x_2-x_1)(x_3-x_1)\ldots(x_n-x_1)\cdot\Delta(x_2, x_3, \ldots, x_n).$$

We do the same to the determinant $\Delta(x_2, x_3, \ldots, x_n)$ on the right-hand side as to $\Delta(x_1, x_2, \ldots, x_n)$. As a result we find that

$$\Delta(x_2, x_3, \ldots, x_n)=(x_3-x_2)\ldots(x_n-x_2)\cdot\Delta(x_3, \ldots, x_n).$$

Continuing similar arguments, we finally find that the original determinant (1.39) is equal to

$$\begin{aligned}\Delta(x_1, x_2, \ldots, x_n)&\\ =(x_2-x_1)(x_3-x_1)\ldots(x_n-x_1)(x_3-x_2)\ldots(x_n-x_2)\ldots&\\ \ldots(x_n-x_{n-1}).&\end{aligned}$$

1.2.6. The determinant of the sum and the product of matrices. It follows directly from the linear property of a determinant that *the determinant of the sum of two square matrices of the same order n, $A=\|a_{ij}\|$ and $B=\|b_{ij}\|$, is equal to the sum of all various determinants of order n which may result if we take a part of the rows (or columns) coinciding with the respective rows (or columns) of the matrix A and the other part coinciding with the respective rows (or columns) of B.*

Let us prove now that *the determinant of the matrix C, which is equal to the product of the square matrix A by the square matrix B, is equal to the product of the determinants of the matrices A and B.*

Let the order of all the three matrices A, B and C be n, and let O be a zero square matrix of order n, and $(-1)\,I$ be the following matrix:

$$(-1)I=\begin{Vmatrix} -1 & 1 & & 0 \\ 0 & -1 & & 0 \\ \cdot & \cdot & & \cdot \\ 0 & 0 & & -1 \end{Vmatrix}.$$

By virtue of Example 2 of 1.2.5, the determinant of the matrix $(-1)I$ is equal to the number $(-1)^n$.

Let us consider the following two square hypermatrices of order $2n$:

$$\begin{Vmatrix} A & O \\ (-1)I & B \end{Vmatrix} \text{ and } \begin{Vmatrix} A & C \\ (-1)I & O \end{Vmatrix}.$$

By virtue of (1.37) and (1.38) of 1.2.5, the determinants of these matrices are

$$\begin{vmatrix} A & O \\ (-1)I & B \end{vmatrix} = |A|\cdot|B|, \quad \begin{vmatrix} A & C \\ (-1)I & O \end{vmatrix} = (-1)^n|(-1)I|\ |C| = (C).$$

Thus, it is sufficient to prove the equality of the determinants

$$\begin{vmatrix} A & O \\ (-1)I & B \end{vmatrix} \text{ and } \begin{vmatrix} A & C \\ (-1)I & O \end{vmatrix}.$$

In more detail, these two determinants can be written as follows:

$$\begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} & 0 & 0 & \dots & 0 \\ a_{21} & a_{22} & & a_{2n} & 0 & 0 & & 0 \\ \cdot & & & & \cdot & & & \cdot \\ a_{n1} & a_{n2} & & a_{nn} & 0 & 0 & & 0 \\ -1 & 0 & & 0 & b_{11} & b_{12} & & b_{1n} \\ 0 & -1 & & 0 & b_{21} & b_{22} & & b_{2n} \\ \cdot & \cdot & \cdot & & \cdot & & & \cdot \\ 0 & 0 & & -1 & b_{n1} & b_{n2} & \dots & b_{nn} \end{vmatrix},$$

$$\begin{vmatrix} a_{11} & a_{12} & & a_{1n} & c_{11} & c_{12} & & c_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} & c_{21} & c_{22} & \dots & c_{2n} \\ \cdot & & & & & & & \\ a_{n1} & a_{n2} & \dots & a_{nn} & c_{n1} & c_{n2} & & c_{nn} \\ -1 & 0 & & 0 & 0 & 0 & & 0 \\ 0 & -1 & & 0 & 0 & 0 & & 0 \\ \cdot & \cdot & \cdot & & \cdot & \cdot & & \cdot \\ 0 & 0 & \dots & -1 & 0 & 0 & & 0 \end{vmatrix} \tag{1.40}$$

To ascertain the equality of these two determinants, it is sufficient to note that the first n columns of these determinants coincide and every column of the second determinant (1.40) labelled $n+k$ (where $k=1, 2, \dots, n$) is obtained, by virtue of the formula $c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}$, by adding to the $(n+k)$th column of the first determinant (1.40) a linear combination of its first n columns with coefficients equal respectively to $b_{k1}, b_{k2}, \dots, b_{kn}$. Thus the determinants (1.40) are equal by virtue of the Corollary 5 of 1.2.3.

It should be pointed out in conclusion that it follows directly from formula (1.37) that the determinant of the direct sum $|A \oplus B| = \begin{vmatrix} A & O \\ O & B \end{vmatrix}$ of two matrices A and B is equal to the product of the determinants of those matrices.

1.2.7. The concept of an inverse matrix. Let A be a square matrix of order n and I, an identity square matrix of the same order (see 1.2).

The matrix B is called a **right inverse** matrix of the matrix A if $AB = I$.

The matrix C is called a **left inverse** matrix of the matrix A if $CA = I$.

Since both matrices A and I are square matrices of order n, the matrices B and C (provided that they exist) are also square matrices of order n.

Let us make sure that *if both matrices B and C exist, then they coincide.* Indeed, on the basis of equations (1.7) (see 1.2), the relations $AB = I$, $CA = I$ and the associative property of the product of matrices, we obtain

$$C = CI = C(AB) = (CA)B = IB = B.$$

Now a question arises about the conditions concerning the matrix A under which both the left and the right inverse matrices of this matrix exist.*

Theorem 1.4. *For the left and right inverse matrices of the matrix A to exist, it is necessary and sufficient that the determinant* $\det A$ *of the matrix A be nonzero.*

Proof. **(1) Necessity.** If at least one of the inverse matrices of the matrix A exists, say, B, then the relation $A \cdot B = I$ yields $\det A \cdot \det B = \det I = 1$**, and hence $\det A \neq 0$.

(2) Sufficiency. Assume that the determinant $\Delta = \det A$ is nonzero. We shall use, as before, the designation A_{ij} for the algebraic adjuncts of the elements a_{ij} of the matrix A and form a matrix B whose ith row contains the algebraic adjuncts of the ith column of the matrix A divided by the value of the determinant Δ:

$$B = \begin{Vmatrix} \frac{A_{11}}{\Delta} & \frac{A_{21}}{\Delta} & \frac{A_{n1}}{\Delta} \\ \frac{A_{12}}{\Delta} & \frac{A_{22}}{\Delta} & \frac{A_{n2}}{\Delta} \\ & & \\ \frac{A_{1n}}{\Delta} & \frac{A_{2n}}{\Delta} & \frac{A_{nn}}{\Delta} \end{Vmatrix}. \tag{1.41}$$

* And obviously those matrices coincide.

** Det $I = 1$ by virtue of Example 2 in 1.2.5.

Let us make sure that this matrix B is both the right and the left inverse matrix of the matrix A.

It is sufficient to prove that the products AB and BA are identity matrices. To do that, it is sufficient to note that in both products *any off-diagonal element is equal to zero* since when the factor $1/\Delta$ is taken outside the matrix this element becomes equal to the sum of the products of the elements of one row (or one column) by the corresponding algebraic adjuncts of another row (or another column). As to the elements lying on the principal diagonal, all those elements in both products AB and BA are equal to unity because the sum of the products of the elements and the corresponding algebraic adjuncts of one row (one column) is equal to the determinant. And thus we have proved the theorem.

Remark 1. The square matrix A whose determinant det A is nonzero is customarily called a **nonsingular** matrix.

Remark 2. In what follows we shall omit the terms "left" and "right" and speak of the *matrix B inverse of the nonsingular matrix A* and defined by the relations $AB=BA=I$. It is also evident that the property of being an inverse matrix is reciprocal in the sense that if B is an inverse of A then A is an inverse of B. We shall use the designation A^{-1} for the inverse of the matrix A.

1.3. Theorem on the Base Minor of a Matrix

1.3.1. The concept of a linear dependence of rows. We have agreed to call the row* $A=(a_1, a_2, \ldots, a_n)$ a linear combination of the rows $B=(b_1, b_2, \ldots, b_n), \ldots, C=(c_1, c_2, \ldots, c_n)$ if for some real numbers $\lambda, \ldots, \mu$ the equations

$$a_j=\lambda b_j+\ldots+\mu c_j \quad (j=1, 2, \ldots, n) \tag{1.42}$$

hold true. It is convenient to write these n equations as one equation

$$A=\lambda B+\ldots+\mu C. \tag{1.43}$$

Every time we come across equation (1.43) we shall understand it in the sense of n equations (1.42).

We shall introduce now the notion of linear dependence of rows.

Definition. *The rows $A=(a_1, a_2, \ldots, a_n)$, $B=(b_1, b_2, \ldots, b_n), \ldots, C=(c_1, c_2, \ldots, c_n)$ are called* ***linearly dependent*** *if there are numbers $\alpha, \beta, \ldots, \gamma$, not all equal to zero and such that the following equalities hold true*:

$$\alpha a_j+\beta b_j+\ldots+\gamma c_j=0 \quad (j=1, 2, \ldots, n). \tag{1.44}$$

* Each row can be regarded as a matrix, and, therefore, it is natural to use capital letters to designate rows.

It is convenient to write n equations (1.44) as one equation

$$\alpha A+\beta B+\ldots+\gamma C=O, \tag{1.45}$$

in which $O=(0, 0, \ldots, 0)$ designates a zero row.

The rows which are not linearly dependent are called **linearly independent.** We can also give "its own" definition of linear independence of rows: *the rows $A, B, \ldots, C$ are called* ***linearly independent*** *if equality* (1.45) *is only possible when all the numbers* $\alpha, \beta, \ldots, \gamma$ are equal to zero.

We shall now prove a simple but very important statement.

Theorem 1.5. *For the rows $A, B, \ldots, C$ to be linearly dependent it is necessary and sufficient that one of these rows be a linear combination of the other rows.*

Proof. **(1) Necessity.** Assume that the rows $A, B, \ldots, C$ are linearly dependent, i. e. equality (1.45) in which at least one of the numbers $\alpha, \beta, \ldots, \gamma$ is nonzero holds true. We assume, for definiteness, that $\alpha\neq 0$. Then, dividing (1.45) by α and introducing the designations $\lambda=-\beta/\alpha, \ldots, \mu=-\gamma/\alpha$, we can rewrite (1.45) in the form

$$A=\lambda B+\ldots+\mu C, \tag{1.46}$$

and this precisely means that the row A is a linear combination of the rows $B, \ldots, C$.

(2) Sufficiency. Assume that one of the rows (say, A) is a linear combination of the other rows. Then, we can find numbers $\lambda, \ldots, \mu$ such that equality (1.46) holds true. But this last equation can be rewritten as

$$(-1)A+\lambda B+\ldots+\mu C=O^*.$$

Since one of the numbers $-1, \lambda, \ldots, \mu$ is nonzero, the last equation establishes a linear dependence of the rows $A, B, \ldots, C$. Thus we have proved the theorem.

As we have agreed, we can replace the term "rows" by the term "columns" in all the arguments presented above.

1.3.2. Theorem on the base minor. Let us consider an arbitrary (not necessarily square) matrix

$$A=\begin{Vmatrix} a_{11} & a_{12} & & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ & & & \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{Vmatrix} \tag{1.47}$$

The kth-order minor of the matrix A is a kth-order determinant with the elements lying in any k rows and any k columns of the matrix A (k is sure not to exceed the least of the numbers m and n).

* Here $O=(0, 0, \ldots, 0)$ is a zero row.

We assume that at least one of the elements a_{ij} of the matrix A is nonzero. Then there is a positive integer r such that the following two conditions are satisfied: (1) the matrix A has a minor of order r different from zero, (2) every minor of the $(r+1)$th or higher order (provided that they exist) is equal to zero.

We call the number r satisfying requirements (1) and (2) the **rank** of the matrix A*. An rth-order minor different from zero is known as a **base minor** (the matrix A can certainly have several rth-order minors different from zero). The rows and columns at whose intersection the base minor is located are called **base rows** and **base columns** respectively.

Let us prove the following fundamental theorem.

Theorem 1.6 (theorem on a base minor). *Base rows (base columns) are linearly independent. Any row (any column) of the matrix A is a linear combination of the base rows (base columns).*

Proof. We present all the arguments for rows.

If base rows were linearly dependent, then, in accordance with Theorem 1.5, one of these rows would be a linear combination of the other base rows and we could subtract the indicated linear combination from this row without changing the value of the base minor and get a row entirely consisting of zeros and this would contradict the fact that the base minor is different from zero. Thus, the base rows are linearly independent.

Let us now prove that any row of the matrix A is a linear combination of the base rows. Since in arbitrary interchanges of rows (or columns) the determinant retains the property of being equal to zero, we can assume, without losing generality, that the base minor is in the upper left corner of matrix (1.47), i. e. lies in the first r rows and the first r columns. Suppose j is any number from 1 to n and k is any number from 1 to m.

Let us ascertain that the determinant of order $r+1$

$$\begin{vmatrix} a_{11} & a_{12} & & a_{1r} & a_{1j} \\ a_{21} & a_{22} & & a_{2r} & a_{2j} \\ \cdot & & & & \\ a_{r1} & a_{r2} & & a_{rr} & a_{rj} \\ a_{k1} & a_{k2} & & a_{kr} & a_{kj} \end{vmatrix} \tag{1.48}$$

is equal to zero. If $j \leqslant r$ or $k \leqslant r$, then the indicated determinant is equal to zero because it has two identical columns or two identical rows.

Now if both numbers j and k exceed r, then (1.48) is a minor of the matrix A of order $(r+1)$ and any such minor is equal to zero (by the

* By definition, the rank of the matrix whose all elements are zeros is equal to zero.

definition of a base minor). Thus, determinant (1.48) is equal to zero for all j from 1 to n and all k from 1 to m.

But then, expanding this determinant according to the last column and designating the cofactors of the elemens of this column independent of the number j as $A_{1j}=c_1$, $A_{2j}=c_2$, ..., $A_{rj}=c_r$, $A_{kj}=c_{r+1}$, we find that

$$c_1a_{1j}+c_2a_{2j}+\ldots+c_ra_{rj}+c_{r+1}a_{kj}=0$$

(for all $j=1, 2, \ldots, n$). Taking into account that in the last equations the cofactor $c_{r+1}=A_{kj}$ coincides with the base minor which is obviously nonzero, we can divide each of these equations by c_{r+1}. But then, introducing the designations

$$\lambda_1=-\frac{c_1}{c_{r+1}},\quad \lambda_2=-\frac{c_2}{c_{r+1}},\ \ldots,\quad \lambda_r=\frac{c_r}{c_{r+1}}$$

we find that $a_{kj}=\lambda_1a_{1j}+\lambda_2a_{2j}+\ldots+\lambda_ra_{rj}$ (for all $j=1, 2, \ldots, n$), and this precisely means that the kth row is a linear combination of the first r (base) rows. And this is what we set out to prove.

1.3.3. The necessary and sufficient condition for a determinant to be zero.

Theorem 1.7. *For the nth-order determinant Δ to be equal to zero, it is necessary and sufficient that its rows (columns) be linearly dependent.*

Proof. **(1) Necessity.** If the nth-order determinant Δ is equal to zero, then the base minor of its matrix has an order r obviously **smaller** than n. But then at least one row is non-base. By Theorem 1.6 that row is a linear combination of the base rows. We can include all the other rows in that linear combination putting zero before them.

Thus, one row is a linear combination of the other rows. But then, by Theorem 1.5, the rows of the determinant are linearly dependent.

(2) Sufficiency. If the rows of Δ are linearly dependent, then, by Theorem 1.5, one row A_i is a linear combination of the other rows. Subtracting the indicated linear combination from the row A_i, we obtain a row consisting entirely of zeros without changing the value of Δ. But then the determinant is equal to zero (by vitue of Corollary 3 from 2.4). We have proved the theorem.

Chapter 2

LINEAR SPACES

The reader is sure to be acquainted from a course of analytic geometry with the operation of addition of free vectors and with the operation of multiplication of a vector by a real number, and also with the properties of these operations. In this chapter, we shall study sets of objects of any nature for whose elements the operation of addition of elements and that of multiplication of an element by a real number are defined in some way (no matter in what way), the indicated operations possessing the same properties as the respective operations on geometric vectors. Sets of this kind are known as linear spaces, they possess a number of common properties which will be established in the present chapter.

2.1. The Concept of a Linear Space

2.1.1. Definition of a linear space. *The set R of elements $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \ldots$ of any nature is called a linear (or affine) space if the following three requirements are met:*

I. *There is a rule according to which any two elements $\boldsymbol{x}$ and $\boldsymbol{y}$ of the set R are associated with a third element $\boldsymbol{z}$ of that set called the* ***sum*** *of the elements $\boldsymbol{x}$ and $\boldsymbol{y}$ and symbolized as $\boldsymbol{z}=\boldsymbol{x}+\boldsymbol{y}$.*

II. *There is a rule according to which any element $\boldsymbol{x}$ of the set R and any real number λ are associated with an element $\boldsymbol{u}$ of that set called the* ***product of the element x by the number*** *λ and symbolized as $\boldsymbol{u}=\lambda\boldsymbol{x}$ or $\boldsymbol{u}=\boldsymbol{x}\lambda$.*

III. *These two rules are subordinated to the following eight axioms.*

1°. $\boldsymbol{x}+\boldsymbol{y}=\boldsymbol{y}+\boldsymbol{x}$ (commutative property of a sum).

2°. $(\boldsymbol{x}+\boldsymbol{y})+\boldsymbol{z}=\boldsymbol{x}+(\boldsymbol{y}+\boldsymbol{z})$ (associative property of a sum).

3°. *There is a zero element* $\mathbf{0}$ *such that* $\boldsymbol{x}+\mathbf{0}=\boldsymbol{x}$ *for any element* $\boldsymbol{x}$ (a special role of a zero element).

4°. *For every element $\boldsymbol{x}$ there is an* ***additive inverse*** *element $\boldsymbol{x}'$ such that $\boldsymbol{x}+\boldsymbol{x}'=\mathbf{0}$.*

5°. $1\cdot\boldsymbol{x}=\boldsymbol{x}$ *for any element* $\boldsymbol{x}$ (a special role of the numerical factor 1).

6°. $\lambda(\mu \boldsymbol{x}) = (\lambda\mu)\boldsymbol{x}$ (associative property with respect to the numerical factor).

7°. $(\lambda+\mu)\boldsymbol{x} = \lambda\boldsymbol{x} + \mu\boldsymbol{x}$ (distributive property with respect to the sum of the numerical factors).

8°. $\lambda(\boldsymbol{x}+\boldsymbol{y}) = \lambda\boldsymbol{x} + \lambda\boldsymbol{y}$ (distributive property with respect to the sum of the elements).

It should be emphasized that when we introduce the notion of a linear space we abstract ourselves not only from the nature of the objects in question but also from the specific form of the rules of formation of a sum of elements and a product of an element by a number (it is only significant that those rules satisfy the eight axioms formulated in the definition given above).

Now if the nature of the objects in question and the form of the rules of formation of a sum of elements and a product of an element by a number are indicated*, then we call such a space a **specific** linear space.

Here are some examples of specific linear spaces.

Example 1. Let us consider a set of all free factors in a three-dimensional space. We define the operations of addition of the indicated vectors and their multiplication by scalars as is usually done in courses of analytic geometry (we define the addition of vectors by the "parallelogram" rule; in the multiplication of a vector by a real number λ, the length of the vector is multiplied by $|\lambda|$ and the direction remains unchanged when $\lambda > 0$, and changes to the opposite when $\lambda < 0$).

It is a simple practice to verify the validity of axioms 1°-8°**.

Thus, the set of all free vectors in a space with the operations of addition and their multiplication by scalars so defined is a linear space which will be designated as B_3.

Similar sets of vectors on a plane and on a straight line, which are also linear spaces, will be denoted by B_2 and B_1 respectively.

Example 2. Let us consider a set $\{x\}$ of all **positive** real numbers. We define the sum of two elements x and y of that set as a product of real numbers x and y (understood in the same sense as in the theory of real numbers). The product of the element x of the set $\{x\}$ by the real number λ is defined as the raising of the real number x to the power λ. The zero element of the set $\{x\}$ is a real number 1 and the additive inverse element (for the given element x) is the real number $1/x$.

The validity of all axioms 1°- 8° is easy to verify. Indeed, the validity of axioms 1° and 2° follows from the commutative and associa-

---

* Needless to say that these rules must be so given that properties 1°-8° presented as axioms in the definition given above be valid.

** The validity of all axioms except for 5°, was established in *Analytic Geometry* by Ilyin and Poznyak and the validity of axiom 5° is doubtless.

tive properties of a product of real numbers; the validity of axioms 3° and 4° follows from the simple equations $x \cdot 1 = x$, $x \cdot \frac{1}{x} = 1$ (for any real $x > 0$), axiom 5° is equivalent to the equation $x^1 = x$, axioms 6° and 7° are valid because for any $x > 0$ and any real λ and μ the relations $(x^\mu)^\lambda = x^{\lambda\mu}$, $x^{\lambda+\mu} = x^\lambda \cdot x^\mu$ hold true, and, finally the validity of axiom 8° follows from the fact that the equality $(xy)^\lambda = x^\lambda y^\lambda$ holds true for any positive x and y and for any real λ. Thus we have ascertained that the set $\{x\}$ with the operations of addition of elements and their multiplication by numbers so defined is a linear space.

Example 3. The set A^n whose elements are ordered collections of n arbitrary real numbers $(x_1, x_2, \ldots, x_n)$ serves as an important example of a linear space. We denote the elements of this set by one symbol $\boldsymbol{x}$, i. e. we write $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ and call the real numbers $x_1, x_2, \ldots, x_n$ the **coordinates** of the element $\boldsymbol{x}$.

In mathematical analysis it is customary to call the set A^n an **n-dimensional coordinate space***. In algebraic interpretation, we can regard A^n as a collection of various rows each of which contains n real numbers (we did just that in 1.3).

The following rules define the operations of addition of the elements of the set A^n and their multiplication by real numbers:

$$(x_1, x_2, \ldots, x_n) + (y_1, y_2, \ldots, y_n) = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n),$$
$$\lambda (x_1, x_2, \ldots, x_n) = (\lambda x_1, \lambda x_2, \ldots, \lambda x_n).$$

We invite the reader to verify the validity of axioms 1°-8° and of the fact that the element $\mathbf{0} = (0, 0, \ldots, 0)$ is a zero element of the set being considered and the element $(-x_1, -x_2, \ldots, -x_n)$ is an additive inverse element of $(x_1, x_2, \ldots, x_n)$.

Example 4. Let us now consider the set $C[a, b]$ of all functions $x = x(t)$ defined and continuous on the interval $a \leqslant t \leqslant b$. The operations of addition of such functions and their multiplication by real numbers are defined by ordinary rules of mathematical analysis. The validity of axioms 1°- 8°, which can be easily verified**, enables us to infer that the set $C[a, b]$ is a linear space.

Example 5. We can consider as an example of a linear space the set $\{P_n(t)\}$ of all algebraic polynomials of a degree not exceeding a natural number n, with the operations defined as in the preceding example. Note that if the set $\{P_n(t)\}$ is considered on the interval $a \leqslant t \leqslant b$, then it is a **subset** of the linear space $C[a, b]$ considered in Example 4.

Remark 1. To explain the notion of a linear space in more detail, we

---

* See *Fundamentals of Mathematical Analysis* by Ilyin and Poznyak, Mir Publishers, Moscow, 1982.

** In particular, function identically equal to zero on the interval $a \leqslant t \leqslant b$ is a zero element of the space $C[a, b]$.

give examples of sets which are not linear spaces for some reason or other:

(a) the set of all vectors of a space except for vectors collinear with some straight line l (since we cannot add together the vectors symmetric about the indicated line l within that set);

(b) the set of all polynomials of a degree exactly equal to the number n (the sum of two polynomials of this kind may turn out to be a polynomial of a degree lower that n);

(c) the set of all polynomials of a degree not exceeding a natural n, whose all coefficients are positive (the elements of such a set cannot be multiplied by negative real numbers).

Remark 2. Note that the elements of an arbitrary linear space are customarily called **vectors.** The fact that the term "vector" is often used in a narrower sense does not lead to a misunderstanding. On the contrary, appealing to familiar geometric notions, we can elucidate and often even foresee a number of results valid for linear spaces of arbitrary nature.

Remark 3. In the definition of a linear space we have formulated we took the numbers λ, μ, . . . from the set of **real** numbers. Therefore, it is natural to call the space we have defined a **real linear space.** In a wider approach, we can take λ, μ, . . . from the set of **complex** numbers. In that case, we arrive at the notion of a **complex linear space.**

2.1.2. Some properties of arbitrary linear spaces. As logical consequences of axioms 1°- 8°, we can obtain a number of statements valid for arbitrary linear spaces. We shall consider two statements as an example.

Theorem 2.1. *There is a single zero element in an arbitrary linear space and there is a single additive inverse element of every element* $\boldsymbol{x}$.

Proof. The existence of **at least one** zero element is stated in axiom 3°. Let us assume that there are two zero elements $\mathbf{0}_1$ and $\mathbf{0}_2$. Then, setting first $\boldsymbol{x}=\mathbf{0}_1$, $\mathbf{0}=\mathbf{0}_2$ in axiom 3° and then $\boldsymbol{x}=\mathbf{0}_2$, $\mathbf{0}=\mathbf{0}_1$, we obtain two equations $\mathbf{0}_1+\mathbf{0}_2=\mathbf{0}_1$, $\mathbf{0}_2+\mathbf{0}_1=\mathbf{0}_2$ whose left-hand sides are equal by virtue of axiom 1°. This means that the right-hand sides of the two last equations are equal because of the transitivity of the equality sign, i. e. $\mathbf{0}_1=\mathbf{0}_2$, and we have thus established the uniqueness of the zero element.

Axiom 4° asserts the existence of **at least one** additive inverse element $\boldsymbol{y}$ of the element $\boldsymbol{x}$. Let us assume that there are two additive inverse elements $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ of a certain element $\boldsymbol{x}$ such that $\boldsymbol{x}+\boldsymbol{y}_1=\mathbf{0}$ and $\boldsymbol{x}+\boldsymbol{y}_2=\mathbf{0}$. But then, by virtue of axioms 3°, 2° and 1°, $\boldsymbol{y}_1=\boldsymbol{y}_1+\mathbf{0}=\boldsymbol{y}_1+(\boldsymbol{x}+\boldsymbol{y}_2)=(\boldsymbol{y}_1+\boldsymbol{x})+\boldsymbol{y}_2=\mathbf{0}+\boldsymbol{y}_2=\boldsymbol{y}_2+\mathbf{0}=\boldsymbol{y}_2$, i.e. $\boldsymbol{y}_1=\boldsymbol{y}_2$, and we have proved the uniqueness of the additive inverse element of any element $\boldsymbol{x}$. We have proved the theorem.

Theorem 2.2. *In an arbitrary linear space*

(1) *the zero element* $\mathbf{0}$ *is equal to the product of the arbitrary element* $\boldsymbol{x}$ *by a real number* 0,

(2) *the additive inverse element of every element* $\boldsymbol{x}$ *is equal to the product of the element* $\boldsymbol{x}$ *by a real number* -1.

Proof. (1) Suppose $\boldsymbol{x}$ is an arbitrary element and $\boldsymbol{y}$ is its additive inverse. Applying successively axioms 3°, 4°, 2°, 5°, 1° and 7° and again 5° and 4°, we find that $\boldsymbol{x}\cdot 0=\boldsymbol{x}\cdot 0+\mathbf{0}=\boldsymbol{x}\cdot 0+(\boldsymbol{x}+\boldsymbol{y})=(\boldsymbol{x}\cdot 0+\boldsymbol{x})+\boldsymbol{y}=(\boldsymbol{x}\cdot 0+\boldsymbol{x}\cdot 1)+\boldsymbol{y}=\boldsymbol{x}(0+1)+\boldsymbol{y}=\boldsymbol{x}\cdot 1+\boldsymbol{y}=\boldsymbol{x}+\boldsymbol{y}=\mathbf{0}$, i.e. $\boldsymbol{x}\cdot 0=\mathbf{0}$.

(2) Suppose $\boldsymbol{x}$ is an arbitrary element and $\boldsymbol{y}=(-1)\cdot x$. Using axioms 5°, 7° and 1° and the equality $\boldsymbol{x}\cdot 0=\mathbf{0}$ we have proved, we get an equality $\boldsymbol{x}+\boldsymbol{y}=\boldsymbol{x}+(-1)\boldsymbol{x}=1\cdot\boldsymbol{x}+(-1)\cdot\boldsymbol{x}=[1+(-1)]\cdot\boldsymbol{x}=0\cdot\boldsymbol{x}=\boldsymbol{x}\cdot 0=\mathbf{0}$ which proves (by virtue of axiom 4°) that $\boldsymbol{y}$ is an additive inverse element of $\boldsymbol{x}$. And that was the proof of the theorem.

We must note in conclusion that axioms 1°-4° make it possible to prove the existence and uniqueness of the **difference** between any two elements $\boldsymbol{x}$ and $\boldsymbol{y}$ of a linear space which is defined as an element $\boldsymbol{z}$ satisfying the condition $\boldsymbol{z}+\boldsymbol{y}=\boldsymbol{x}$*. (The sum $\boldsymbol{z}=\boldsymbol{x}+(-1)\boldsymbol{y}$ is precisely that element.)

2.2. The Basis and the Dimensionality of a Linear Space

2.2.1. The concept of a linear dependence of the elements of a linear space. In the course of analytic geometry** we introduced the notion of a linear dependence of vectors, and in 1.3.1 of the present book we introduced the notion of a linear dependence of rows (or, what is the same thing, of the elements of the space A^n considered in Example 3 in 2.1.1).

A generalization of these notions is the notion of linear dependence of elements of an arbitrary linear space which we shall now consider.

Let us consider an arbitrary real linear space R with elements $\boldsymbol{x}, \boldsymbol{y}, \ldots, \boldsymbol{z}, \ldots$.

A ***linear combination*** of the elements $\boldsymbol{x}, \boldsymbol{y}, \ldots, \boldsymbol{z}$ *of the space* R *is the sum of the products of these elements by arbitrary real numbers*, i.e. an expression of the form

$$\alpha\boldsymbol{x}+\beta\boldsymbol{y}+\ldots+\gamma\boldsymbol{z}, \tag{2.1}$$

where $\alpha, \beta, \ldots, \gamma$ are any real numbers.

Definition 1. The elements $\boldsymbol{x}, \boldsymbol{y}, \ldots, \boldsymbol{z}$ *of the space* R *are said to be* ***linearly dependent*** *if there are real numbers* $\alpha, \beta, \ldots, \gamma$ *at least one of which is different from zero, such that the linear combination of*

---

* It is sufficent to repeat the proof given in the theory of real numbers in *Fundamentals of Mathematical Analysis*, Ilyin and Poznyak.

** See *Analytic Geometry*, Ilyin and Poznyak, 2.1.3.

the elements $\boldsymbol{x}, \boldsymbol{y}, \ldots, \boldsymbol{z}$ *with the indicated numbers is a zero element of the space* R, *i.e. there holds an equality*

$$\alpha \boldsymbol{x} + \beta \boldsymbol{y} + \ldots + \gamma \boldsymbol{z} = 0. \tag{2.2}$$

The elements $\boldsymbol{x}, \boldsymbol{y}, \ldots, \boldsymbol{z}$ which are not linearly dependent are said to be **linearly independent.**

Here is another definition of linearly independent vectors based on the logical negation of content of Definition 1.

Definition 2. *The elements* $\boldsymbol{x}, \boldsymbol{y}, \ldots, \boldsymbol{z}$ *of the space* R *are said to be* ***linearly independent*** *if the linear combination* (2.1) *is a zero element of the space* R *only under the condition* $\alpha = \beta = \ldots = \gamma = 0$.

Theorem 2.3. *For the elements* $\boldsymbol{x}, \boldsymbol{y}, \ldots, \boldsymbol{z}$ *of the space* R *to be linearly dependent, it is necessary and sufficient that one of these elements be a linear combination of the other elements.*

Proof. **(1) Necessity.** Suppose the elements $\boldsymbol{x}, \boldsymbol{y}, \ldots, \boldsymbol{z}$ are linearly dependent, i.e. equality (2.2), in which at least one of the numbers $\alpha, \beta, \ldots, \gamma$ is nonzero, holds true. Assume, for definiteness, $\alpha \neq 0$. Then, dividing (2.2) by α and introducing the designations $\lambda = -\frac{\beta}{\alpha}, \ldots, \mu = -\frac{\gamma}{\alpha}$, we can rewrite (2.2) in the form

$$\boldsymbol{x} = \lambda \boldsymbol{y} + \ldots + \mu \boldsymbol{z}, \tag{2.3}$$

and this means that the element $\boldsymbol{x}$ is a linear combination of the elements $\boldsymbol{y}, \ldots, \boldsymbol{z}$.

(2) Sufficiency. Assume that one of the elements (say, $\boldsymbol{x}$) is a linear combination of the other elements. Then, there can be found numbers $\lambda, \ldots, \mu$ such that equality (2.3) holds true. But this last equality can be rewritten as

$$(-1)\, \boldsymbol{x} + \lambda \boldsymbol{y} + \ldots + \mu \boldsymbol{z} = \boldsymbol{0}. \tag{2.4}$$

Since one of the numbers $(-1), \lambda, \ldots, \mu$ is different from zero, equation (2.4) establishes a linear dependence of the elements $\boldsymbol{x}, \boldsymbol{y}, \ldots, \boldsymbol{z}$. And that is what we had to prove.

The following two simple statements hold true.

1. *If there is a zero element among the elements* $\boldsymbol{x}, \boldsymbol{y}, \ldots, \boldsymbol{z}$, *then these elements are linearly dependent.* Indeed, if, say, $\boldsymbol{x} = \boldsymbol{0}$, then equality (2.2) is valid for $\alpha = 1$, $\beta = \ldots = \gamma = 0$.

2. *If a part of the elements* $\boldsymbol{x}, \boldsymbol{y}, \ldots, \boldsymbol{z}$ *are linearly dependent, then all these elements are linearly dependent.* Indeed, if, say, the elements $\boldsymbol{y}, \ldots, \boldsymbol{z}$ are linearly dependent, then the equality $\beta \boldsymbol{y} + \ldots + \gamma \boldsymbol{z} = \boldsymbol{0}$, in which not all the numbers $\beta, \ldots, \gamma$ are equal to zero, holds true. But then equality (2.2) with the same numbers $\beta, \ldots, \gamma$ and with $\alpha = 0$ holds true.

We conclude with studying the linear dependence of the elements of the space A^n introduced in Example 3 in 2.1.1.

We shall prove that n elements of the indicated space

$$\left.\begin{aligned} e_1&=(1,\ 0,\ 0,\ \ldots,\ 0),\\ e_2&=(0,\ 1,\ 0,\ \ldots,\ 0),\\ e_3&=(0,\ 0,\ 1,\ \ldots,\ 0),\\ &\ldots\\ e_n&=(0,\ 0,\ 0,\ldots,\ 1)\end{aligned}\right\} \tag{2.5}$$

are linearly independent and the collection of n elements (2.5) and one more **arbitrary** element $\boldsymbol{x}=(x_1,\ x_2,\ \ldots,\ x_n)$ of the space A^n forms a linearly dependent system of elements.

Let us consider the linear combination of elements (2.5) with any numbers $\alpha_1,\ \alpha_2,\ \alpha_3,\ \ldots,\ \alpha_n$. By virtue of the axioms, this linear combination is an element

$$\alpha_1\boldsymbol{e}_1+\alpha_2\boldsymbol{e}_2+\ldots+\alpha_n\boldsymbol{e}_n=(\alpha_1,\ \alpha_2,\ \ldots,\ \alpha_n),$$

which is a zero element only under the condition $\alpha_1=\alpha_2=\ldots=\alpha_n=0$. But this precisely means a linear independence of elements (2.5).

Let us now prove that a system consisting of n elements (2.5) and one more **arbitrary** element $\boldsymbol{x}=(x_1,\ x_2,\ \ldots,\ x_n)$ of the space A^n is linearly dependent. By virtue of Theorem 2.3, it is sufficient to prove that the *element* $\boldsymbol{x}=(x_1,\ x_2,\ \ldots,\ x_n)$ *is a linear combination of elements* (2.5) and this is obvious since by virtue of the axioms

$$\boldsymbol{x}=(x_1,\ x_2,\ \ldots,\ x_n)=x_1\boldsymbol{e}_1+x_2\boldsymbol{e}_2+\ldots+x_n\boldsymbol{e}_n.$$

2.2.2. The basis and the coordinates. Let us consider an arbitrary real linear space R.

Definition. *A collection of linearly independent elements* $\boldsymbol{e}_1,\ \boldsymbol{e}_2,\ \ldots,\ \boldsymbol{e}_n$ *of the space* R *is said to be the* ***basis*** *of that space if for every element* $\boldsymbol{x}$ *of the space* R *there are real numbers* $x_1,\ x_2,\ \ldots,\ x_n$ *such that the following equality holds true:*

$$\boldsymbol{x}=x_1\boldsymbol{e}_1+x_2\boldsymbol{e}_2+\ldots+x_n\boldsymbol{e}_n. \tag{2.6}$$

In that case, equation (2.6) is called the **resolution of the element** $\boldsymbol{x}$ *into components* $\boldsymbol{e}_1,\ \boldsymbol{e}_2,\ \ldots,\ \boldsymbol{e}_n$ and the numbers $x_1,\ x_2,\ \ldots,\ x_n$ are called the **coordinates** of the element $\boldsymbol{x}$ (with respect to the basis $\boldsymbol{e}_1,\ \boldsymbol{e}_2,\ \ldots,\ \boldsymbol{e}_n$).

Let us prove that *every element* $\boldsymbol{x}$ *of the linear space* R *can be resolved into the components* $\boldsymbol{e}_1,\ \boldsymbol{e}_2,\ \ldots,\ \boldsymbol{e}_n$ ***in only one way,*** i.e. the coordinates of every element $\boldsymbol{x}$ with respect to the basis $\boldsymbol{e}_1,\ \boldsymbol{e}_2,\ \ldots,\ \boldsymbol{e}_n$ are defined **uniquely.**

Let us assume that besides resolution (2.6) one more resolution into the same components

$$\boldsymbol{x}=x_1'\boldsymbol{e}_1+x_2'\boldsymbol{e}_2+\ldots+x_n'\boldsymbol{e}_n \tag{2.7}$$

is possible for a certain element $\boldsymbol{x}$.

A term-by-term subtraction of equations (2.6) and (2.7) leads to a relation*

$$(x_1-x_1')\boldsymbol{e}_1+(x_2-x_2')\boldsymbol{e}_2+\ldots+(x_n-x_n')\boldsymbol{e}_n=\boldsymbol{0}. \tag{2.8}$$

Because of the linear independence of the base elements $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$, relation (2.8) leads to equations $x_1-x_1'=0$, $x_2-x_2'=0$, $\ldots, x_n-x_n'=0$ or $x_1=x_1'$, $x_2=x_2'$, $\ldots$, $x_n=x_n'$. We have proved the uniqueness of resolution into components. The practical significance of the basis lies also in the fact that when it is specified the operations of addition of elements and their multiplication by numbers become the corresponding operations on numbers, the coordinates of those elements. The following assertion holds true.

Theorem 2.4 *When any two elements of the linear space R are added together, their coordinates (with respect to any basis of the space R) are added; when an arbitrary element is multiplied by any number λ, all the coordinates of that element are multiplied by λ.*

Proof. Suppose $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ is an arbitrary basis of the space R, $\boldsymbol{x}=x_1\boldsymbol{e}_1+x_2\boldsymbol{e}_2+\ldots+x_n\boldsymbol{e}_n$ and $\boldsymbol{y}=y_1\boldsymbol{e}_1+y_2\boldsymbol{e}_2+\ldots+y_n\boldsymbol{e}_n$ are any two elements of that space.

Then, by virtue of axioms 1°-8°, we have

$$\boldsymbol{x}+\boldsymbol{y}=(x_1+y_1)\boldsymbol{e}_1+(x_2+y_2)\boldsymbol{e}_2+\ldots+(x_n+y_n)\boldsymbol{e}_n,$$
$$\lambda\boldsymbol{x}=(\lambda x_1)\boldsymbol{e}_1+(\lambda x_2)\boldsymbol{e}_2+\ldots+(\lambda x_n)\boldsymbol{e}_n.$$

This completes the proof since the resolution into components is unique.

Here are some examples of the bases of specific linear spaces.

It is known from analytic geometry that any three noncomplanar vectors form a basis in the linear space B_3 of all free vectors (this space was considered in example 1 in 2.1.1).

Now note that the collection of n elements (2.5) considered at the end of 2.2.1 forms a basis in the linear space A^n introduced in Example 3 in 2.1.1.

In fact, we have proved at the end of 2.2.1 that elements (2.5) are linearly independent and that any element $\boldsymbol{x}=(x_1, x_2, \ldots, x_n)$ of the space A^n is a linear combination of elements (2.5).

Let us finally ascertain that the basis of the linear space $\{x\}$ introduced in Example 2 in 2.1.1, consists of one element which can be any **nonzero** element of that space (i.e. any positive real number x_0 not equal to 1). It is sufficient to prove that for any positive real number x there is a real number λ such that $x=x_0^{\lambda}$**. But this is obvious: it is sufficient to take $\lambda=\log_{x_0}x$.

---

* The possibility of a term-by-term subtraction of equations (2.6) and (2.7) and the grouping of the terms we perform follows axioms 1°-8°.

** Recall that the product of the element x_0 by the number λ is defined as the number x_0^{λ}.

2.2.3. The dimensionality of a linear space. As before, we shall consider an arbitrary real linear space R.

Definition 1. *The linear space R is called* ***n-dimensional*** *if there are n linearly independent elements in it whereas any $(n+1)$ elements are linearly dependent.* The number n is called a **dimension** of the space R.

The dimension of the space R is usually designated as $\dim R$.

Definition 2. *The linear space R is said to be* ***infinite-dimensional*** *if it contains any number of linearly independent elements.**

In this book we shall study only spaces of finite dimension n. Infinite-dimensional spaces form a subject requiring special investigation. (They are studied in chapters 10 and 11 of *Fundamentals of Mathematical Analysis*, Part 2.)

Let us elucidate the connection between the notion of dimensionality of a space and that of a basis introduced in 2.2.2.

Theorem 2.5. *If R is a linear space of dimension n, then any n linearly independent elements of that space form its basis.*

Proof. Suppose $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ is any system of n linearly independent elements of the space R (the existence of at least one system of that kind follows from Definition 1).

If $\boldsymbol{x}$ is any element of R, then, according to Definition 1, a system of $(n+1)$ elements $\boldsymbol{x}, \boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ is linearly dependent, i.e. there are numbers $\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_n$ which are not all equal to zero and such that the equality

$$\alpha_0\boldsymbol{x}+\alpha_1\boldsymbol{e}_1+\alpha_2\boldsymbol{e}_2+\ldots+\alpha_n\boldsymbol{e}_n=0 \tag{2.9}$$

holds true. Note that the number α_0 is obviously nonzero (since otherwise equation (2.9) would yield a linear dependence of the elements $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$). But then, dividing equation (2.9) by α_0 and setting $x_1=-\frac{\alpha_1}{\alpha_0}, x_2=-\frac{\alpha_2}{\alpha_0}, \ldots, x_n=-\frac{\alpha_n}{\alpha_0}$, we get

$$\boldsymbol{x}=x_1\boldsymbol{e}_1+x_2\boldsymbol{e}_2+\ldots+x_n\boldsymbol{e}_n \tag{2.10}$$

from (2.9). Since $\boldsymbol{x}$ is an arbitrary element of R, equation (2.10) proves that the system of elements $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ is a basis of the space R. We have proved the theorem.

Theorem 2.6. *If the linear space R has a basis consisting of n elements, then the dimension of R is n.*

Proof. Suppose the system of n elements $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ is a basis of the space R. It is sufficient to prove that **any** $(n+1)$ elements $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n+1}$ of that space are linearly dependent.** Resolv-

---

* The fact that the space R is infinite-dimensional is symbolized as $\dim R=\infty$.

** Because the basis elements $e_1, e_2, \ldots, e_n$ form a system of n linearly independent elements of the space R.

ing each of these elements into components, we get

$$\begin{aligned}
x_1 &= a_{11}e_1 + a_{12}e_2 + \ldots + a_{1n}e_n,\\
x_2 &= a_{21}e_1 + a_{22}e_2 + \ldots a_{2n}e_n,\\
&\ldots\ldots\ldots\\
x_{n+1} &= a_{(n+1)1}e_1 + a_{(n+1)2}e_2 + \ldots + a_{(n+1)n}e_n,
\end{aligned}$$

where $a_{11}, a_{12}, \ldots, a_{(n+1)n}$ are some real numbers.

The linear dependence of the elements $x_1, x_2, \ldots, x_{n+1}$ is evidently equivalent to the linear dependence of the rows of the matrix

$$A = \begin{Vmatrix} a_{11} & a_{12} & & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{(n+1)1} & a_{(n+1)2} & & a_{(n+1)n} \end{Vmatrix}$$

But the rows of this matrix are obviously linearly dependent since the order of the base minor of this matrix (containing $n+1$ rows and n columns) does not exceed n and at least one of its $n+1$ rows is not a base row and is, by the theorem on a base minor*, a linear combination of base (and, evidently, all the other) rows. We have proved the theorem.

Directing our attention to the examples considered at the end of 2.2.2, we can say that the dimension of the space B_3 of all free vectors is 3, that of the space A^n is n and the dimension of the space $\{x\}$ is unity.

The linear space $C[a, b]$ of all functions $x = x(t)$, defined and continuous on the interval $a \leqslant t \leqslant b$, can serve as an example of an infinite-dimensional space (see Example 4 in 2.1.1).

Indeed, for any number n, a system of $(n+1)$ elements $1, t, t^2, \ldots, t^n$ of that space is linearly independent (since otherwise a certain polynomial $C_0 + C_1 t + C_2 t^2 + \ldots + C_n t^n$, in which not all the coefficients $C_0, C_1, \ldots, C_n$ are equal to zero, would be identically equal to zero on the interval $a \leqslant t \leqslant b$).

2.2.4. The concept of isomorphism of linear spaces. We show here that various linear spaces of one and the same dimension n, in the sense of the properties connected with the operations introduced in those spaces, do not differ essentially from one another.

Since we have introduced in linear spaces only the operations of addition of elements and their multiplication by numbers, it is natural to formulate the following definition.

Definition. *Two arbitrary real linear spaces R and R' are said to be* ***isomorphic*** *(or, more precisely, linearly* ***isomorphic****)*

* See Theorem 1.6 in 1.3.2.

to each other if a one-to-one correspondence can be established between the elements of those spaces so that if the elements* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *of the space* R *are associated with the respective elements* $\boldsymbol{x}'$ *and* $\boldsymbol{y}'$ *of the space* R', *then the element* $\boldsymbol{x}+\boldsymbol{y}$ *is associated with the element* $\boldsymbol{x}'+\boldsymbol{y}'$ *and the element* $\lambda\boldsymbol{x}$ *is associated with the element* $\lambda\boldsymbol{x}'$ *for any real* λ.

Note that *if the linear space* R *and* R' *are isomorphic to each other, then a zero element of* R *is associated with a zero element of* R' *and vice versa.* (Indeed, suppose the element $\boldsymbol{x}$ of R is associated with a certain element $\boldsymbol{x}'$ of R'. Then the element $0\cdot\boldsymbol{x}$ of R is associated with the element $0\cdot\boldsymbol{x}'$ of R'.)

It follows that if in the case of isomorphism the elements $\boldsymbol{x}, \boldsymbol{y}, \ldots, \boldsymbol{z}$ of the space R correspond to the respective elements $\boldsymbol{x}', \boldsymbol{y}', \ldots, \boldsymbol{z}'$ of the space R', then the linear combination $\alpha\boldsymbol{x}+\beta\boldsymbol{y}+\ldots+\gamma\boldsymbol{z}$ is a zero element of the space R if and only if the linear combination $\alpha\boldsymbol{x}'+\beta\boldsymbol{y}'\ldots+\gamma\boldsymbol{z}'$ is a zero element of the space R'.

But this means that *if the spaces* R *and* R' *are isomorphic, then the maximum number of linearly independent elements in each of these spaces is the same.*

In other words, *two isomorphic spaces must have the same dimension.*

Hence, *spaces of different dimensions cannot be isomorphic.*

Let us now prove the following statement.

Theorem 2.7. *Any two n-dimensional real linear spaces* R *and* R' *are isomorphic.*

Proof. Let us choose some basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ in R and some basis $\boldsymbol{e}'_1, \boldsymbol{e}'_2 \ldots, \boldsymbol{e}'_n$ in R'. We put every element $\boldsymbol{x}'=x_1\boldsymbol{e}'_1+x_2\boldsymbol{e}'_2+\ldots+x_n\boldsymbol{e}'_n$ of the space R' in correspondence with the element $\boldsymbol{x}=x_1\boldsymbol{e}_1+x_2\boldsymbol{e}_2+\ldots+x_n\boldsymbol{e}_n$ of the space R (i.e. we take as $\boldsymbol{x}'$ the element of R' which has *the same coordinates* with respect to the basis $\boldsymbol{e}'_1, \boldsymbol{e}'_2, \ldots, \boldsymbol{e}'_n$ as the element $\boldsymbol{x}$ with respect to the basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$).

We shall verify that the correspondence we have established is one-to-one. Indeed, every element $\boldsymbol{x}$ of the space R is in a one-to-one correspondence with the coordinates $x_1, x_2, \ldots, x_n$, which, in their turn, define a single element $\boldsymbol{x}'$ of the space R'. Because of the equivalence of the spaces R and R', every element $\boldsymbol{x}'$ of the space R' is associated with only one element $\boldsymbol{x}$ of the space R.

It remains to point out that if the elements $\boldsymbol{x}$ and $\boldsymbol{y}$ of R are put into correspondence with the respective elements $\boldsymbol{x}'$ and $\boldsymbol{y}'$ of R', then, by virtue of Theorem 2.4, the element $\boldsymbol{x}+\boldsymbol{y}$ is associated

* Recall that the correspondence between the elements of two sets R and R' is said to be **one-to-one** if that correspondence relates every element of R to one and only one element of R'. In that case every element of R' corresponds to one and only one element of R.

with the element $x'+y'$ and the element λx with the element $\lambda x'$. We have proved the theorem.

It follows from the consideration presented above that the only essential characteristic of a finite-dimensional linear space is its dimensionality.

2.3. Subspaces of Linear Spaces

2.3.1. The concepts of a subspace and a linear closure. Let us assume that a subset L of the linear space R meets the following **two requirements.**

1°. *If the elements x and y belong to the subset L, then the sum $x+y$ also belongs to that subset.*

2°. *If the element x belongs to the subset L and λ is any real number, then the element λx also belongs to the subset L.*

Let us make sure that the *subset L meeting requirements* 1° *and* 2° *is itself a linear space.* It is sufficient to verify the validity of axioms 1°-8° from the definition of a linear space for the elements of the subset L. All the indicated axioms, except for 3° and 4° are obviously valid for the elements of the subset L since they are valid for **all** elements of the space R. It remains to verify the validity of axioms 3° and 4°. Suppose x is any element of the subset L and λ is any real number. Then, by virtue of requirement 2°, the element λx also belongs to L. It remains to note that (according to Theorem 2.2) for $\lambda=0$ this element λx turns into a zero element of the space R and for $\lambda=-1$ it turns into an additive inverse element of x. Thus, the zero element and the additive inverse element (of every x) belong to the subset L and this means that axioms 3° and 4° are valid for the elements of the subset L. We have thus completely proved that the subset L is itself a linear space.

Definition. *The subset L of the linear space R meeting requirements* 1° *and* 2° *is called a* ***linear subspace*** *(or simply* ***subspace****) of the space R.*

Here are some examples of subspaces:

(1) the so-called **null subspace,** i.e. a subset of the linear space R consisting of one zero element,

(2) the whole space R (which can certainly be regarded as a subspace).

Both these subspaces are customarily called **improper.**

Here are some more interesting examples of subspaces.

Example 1. The subset $\{P_n(t)\}$ of all algebraic polynomials of a degree not exceeding a natural number n*, in the linear space $C[a, b]$ of all functions $x=x(t)$ defined and continuous on the in-

* This subset was introduced in Example 5 in 2.1.1.

terval $a \leqslant t \leqslant b$* (the validity of requirements 1° and 2° for the elements of the subset $\{P_n(t)\}$ is obvious).

Example 2. The subset B_2 of all free vectors, parallel to some plane, in the linear space B_3 of all free vectors** (the validity of requirements 1° and 2° for the elements of B_2 is obvious).

Example 3. Assume that $\boldsymbol{x}, \boldsymbol{y}, \ldots, \boldsymbol{z}$ is a collection of elements of a certain linear space R. *The* ***linear closure*** *of the elements* $\boldsymbol{x}, \boldsymbol{y}, \ldots, \boldsymbol{z}$ *is a collection of all linear combinations of those elements*, i.e. the set of elements of the form

$$\alpha \boldsymbol{x} + \beta \boldsymbol{y} + \ldots + \gamma \boldsymbol{z},$$

where $\alpha, \beta, \ldots, \gamma$ are any real numbers.

We shall agree to use the symbol $L(\boldsymbol{x}, \boldsymbol{y}, \ldots, \boldsymbol{z})$ for the linear closure of the elements $\boldsymbol{x}, \boldsymbol{y}, \ldots, \boldsymbol{z}$.

Requirements 1° and 2° formulated at the beginning of this section are evidently satisfied for the linear closure of the arbitrary elements $\boldsymbol{x}, \boldsymbol{y}, \ldots, \boldsymbol{z}$ of the linear space R. Therefore, *every linear closure is a subspace of the principal linear space* R. This subspace evidently contains the elements $\boldsymbol{x}, \boldsymbol{y}, \ldots, \boldsymbol{z}$ on which the linear closure $L(\boldsymbol{x}, \boldsymbol{y}, \ldots, \boldsymbol{z})$ is constructed. On the other hand, every subspace containing the elements $\boldsymbol{x}, \boldsymbol{y}, \ldots, \boldsymbol{z}$ must also contain all linear combinations of those elements. Therefore, *the linear closure of the elements* $\boldsymbol{x}, \boldsymbol{y}, \ldots, \boldsymbol{z}$ *is the* ***least*** *subspace containing the elements* $\boldsymbol{x}, \boldsymbol{y}, \ldots, \boldsymbol{z}$.

The linear closure of the elements $1, t, t^2, \ldots, t^n$ of the linear space $C[a, b]$ of all functions $x = x(t)$ defined and continuous on the interval $a \leqslant t \leqslant b$ can serve as an example of a linear closure. This linear closure is evidently a set $\{P_n(t)\}$ of all algebraic polynomials of a degree not exceeding n.

Some other examples of subspaces will be studied in 2.3.3.

Let us now consider the dimensionality of a subspace (of a linear closure, in particular).

We can state that *the dimension of any subspace of an n-dimensional linear space R does not exceed the dimension n of the space R* (since every linearly independent system of elements of a subspace is at the same time a linearly independent system of elements of the entire space R).

To be more precise, we can state that *if the subspace L does not coincide with the whole n-dimensional linear space R, then the dimension of L is strictly less than n.*

This follows from the fact that if the dimensions of L and R are both equal to n, then every basis of the subspace L is (by virtue of Theorem 2.5) the basis of the whole space R since it consists of n elements.

Note that if a basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ has been chosen in the whole space

*The space $C[a,b]$ was introduced in Example 4 in 2.1.1.

** The sets B_2 and B_3 were introduced in Example 1 in 2.1.1.

R, then the base elements of the subspace L cannot, in general, be chosen from the elements $e_1, e_2, \ldots, e_n$ (since, in a general case, none of the elements $e_1, e_2, \ldots, e_n$ necessarily belong to L). The converse **statement** is true, however: *if the elements $e_1, e_2, \ldots, e_n$ form the basis of a k-dimensional subspace of an n-dimensional linear space R, then this basis can be complemented by the elements $e_{k+1}, \ldots, e_n$ of the space R so that the collection of the elements $e_1, \ldots, e_k, e_{k+1}, \ldots$ $\ldots, e_n$ forms a basis of the whole space R.*

Let us prove this statement. If $k<n$, then there is an element e_{k+1} of the space R such that the elements $e_1, e_2, \ldots, e_k, e_{k+1}$ are linearly independent (otherwise, the space R would be k-dimensional). Furthermore, if $k+1<n$, then there is an element e_{k+2} of the space R such that the elements $e_1, e_2, \ldots, e_k, e_{k+1}, e_{k+2}$ are linearly independent (otherwise, the space R would be $(k+1)$-dimensional). Continuing a similar argument, we shall prove the statement we have formulated.

In conclusion, we shall prove a significant theorem on the dimensionality of a linear closure.

Theorem 2.8. *The dimension of the linear closure $L(x, y, \ldots, z)$ of the elements $x, y, \ldots, z$ is equal to maximum number of linearly independent elements in the system of elements $x, y, \ldots, z$. In particular, if the elements $x, y, \ldots, z$ are linearly independent, then the dimension of the linear closure $L(x, y, \ldots, z)$ is equal to the number of elements $x, y, \ldots, z$ (and these elements themselves form a basis of the linear closure $L(x, y, \ldots, z)$).*

Proof. We assume that there are r linearly independent elements among the elements $x, y, \ldots, z$ (we designate them as $x_1, x_2, \ldots, x_r$), and any $(r+1)$ elements from the elements $x, y, \ldots, z$ are linearly dependent. Then each element $x, y, \ldots, z$ is a certain linear combination of the elements $x_1, x_2, \ldots, x_r$*, and, since, by definition, every element of the linear closure $L(x, y, \ldots, z)$ is a certain linear combination of elements $x, y, \ldots, z$ every element of the indicated linear closure is a certain linear combination of the elements $x_1, x_2, \ldots, x_r$ alone. But this precisely means that the system of linearly independent elements $x_1, x_2, \ldots, x_r$ forms a basis of the linear closure $L(x, y, \ldots, z)$ and that the dimensionality of $L(x, y, \ldots, z)$ is equal to r. Thus the proof is complete.

2.3.2. A new definition of a rank of a matrix. In 1.3 we defined the rank of the arbitrary matrix A as the **order of its base minor,** i.e. as the number r satisfying the requirement that the matrix A should have a nonzero minor of order r and should not have nonzero minors of order exceeding r.

We shall prove in this subsection that *the rank of the arbitrary*

---

* This fact can be established with the aid of the same arguments which were presented in the proof of Theorem 2.5.

matrix A is equal to the maximum number of linearly independent rows (or columns) of that matrix.

This statement yields a new definition of the rank of a matrix as a maximum number of linearly independent rows (or columns) of that matrix*.

We present all the arguments for the rows (since they are similar for the columns). We consider the linear closure of the base rows of an arbitrary matrix A, containing m rows and n columns, in the linear space A^n (introduced in Example 3 in 2.1.1) and assume that the number of base rows is equal to r. It follows from Theorem 1.6 on a base minor that any row of the matrix A is an element of the indicated linear closure, and from the linear independence of r base rows and from Theorem 2.8 it follows that the dimension of the indicated linear closure is equal to r. Hence any $(r+1)$ elements of that linear closure (and, in particular, any $(r+1)$ rows of the matrix A) are linearly dependent. And this means that the number r is the maximum number of linearly independent rows.

2.3.3. The sum and the intersection of subspaces. Suppose L_1 and L_2 are two arbitrary subspaces of one and the same linear space R.

The collection of all elements x of the space R belonging simultaneously to L_1 and L_2 forms a subspace of the space R** called the **intersection** of the subspaces L_1 and L_2.

The collection of all elements of the space R of the form $\boldsymbol{y}+\boldsymbol{z}$, where $\boldsymbol{y}$ is an element of the subspace L_1 and $\boldsymbol{z}$ is an element of the subspace L_2, forms a subspace of the space R*** called the **sum** of the subspaces L_1 and L_2.

Example. Suppose R is a linear space of all free vectors (in a three-dimensional space), L_1 is a subspace of all free vectors which are parallel to the plane Oxy and L_2 is a subspace of all free vectors which are parallel to the plane Oxz. Then the sum of the subspaces L_1 and L_2 is the whole space R**** and the intersection of the subspaces L_1 and L_2 is the set of all free vectors which are parallel to the axis Ox.

The following statement holds true.

Theorem 2.9. *The sum of the dimensions of the arbitrary subspaces L_1 and L_2 of the finite-dimensional linear space R is equal to*

---

* In particular, this statement yields a rather nontrivial theorem stating that in any matrix the maximum number of linearly independent rows coincides with the maximum number of linearly independent columns.

** Since the elements of this collection meet requirements 1° and 2° formulated at the beginning of 2.3.1.

*** See the preceding footnote.

**** Indeed, any vector $\boldsymbol{x}$ of the space R is a linear combination $\boldsymbol{x}=\alpha\boldsymbol{i}+\beta\boldsymbol{j}+\gamma\boldsymbol{k}$ of the base vectors $\boldsymbol{i}$, $\boldsymbol{j}$, $\boldsymbol{k}$, which are parallel to the axes Ox, Oy and Oz respectively, the vector $\alpha\boldsymbol{i}+\beta\boldsymbol{j}$ belonging to L_1 and the vector $\gamma\boldsymbol{k}$ to L_2.

the sum of the dimension of the intersection of these subspaces and the dimension of the sum of these subspaces.

Proof. Let us designate the intersection of L_1 and L_2 as L_0 and the sum of L_1 and L_2 as $\widehat{L}$. Considering L_0 to be k-dimensional, we choose in it a basis

$$\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_k. \tag{2.11}$$

Using the assertion proved in 2.3.1, we complete the basis (2.11) to form a basis

$$\boldsymbol{e}_1, \ldots, \boldsymbol{e}_k, \quad \boldsymbol{g}_1, \ldots, \boldsymbol{g}_l \tag{2.12}$$

in the subspace L_1, and to form a basis

$$\boldsymbol{e}_1, \ldots, \boldsymbol{e}_k, \quad \boldsymbol{f}_1, \ldots, \boldsymbol{f}_m \tag{2.13}$$

in the subspace L_2.

It is sufficient to prove that the elements

$$\boldsymbol{g}_1, \ldots, \boldsymbol{g}_l, \quad \boldsymbol{e}_1, \ldots, \boldsymbol{e}_k, \quad \boldsymbol{f}_1, \ldots, \boldsymbol{f}_m \tag{2.14}$$

are the basis of the sum $\widehat{L}$ of the subspaces L_1 and L_2*. To do that, it is, in turn, sufficient to prove that elements (2.14) are linearly independent and that any element $\boldsymbol{x}$ of the sum $\widehat{L}$ is a certain linear combination of elements (2.14).

We shall first prove that elements (2.14) are linearly independent.

Let us assume that a certain linear combination of elements (2.14) is a zero element, i.e. the equality

$$\alpha_1\boldsymbol{g}_1+\ldots+\alpha_l\boldsymbol{g}_l+\beta_1\boldsymbol{e}_1+\ldots+\beta_k\boldsymbol{e}_k+\gamma_1\boldsymbol{f}_1+\ldots+\gamma_m\boldsymbol{f}_m=0 \tag{2.15}$$

or

$$\alpha_1\boldsymbol{g}_1+\ldots+\alpha_l\boldsymbol{g}_l+\beta_1\boldsymbol{e}_1+\ldots+\beta_k\boldsymbol{e}_k=-\gamma_1\boldsymbol{f}_1-\ldots-\gamma_m\boldsymbol{f}_m \tag{2.16}$$

holds true. Since the left-hand side of (2.16) is an element of L_1 and the right-hand side of (2.16) is an element of L_2, it follows that both the left-hand and right-hand sides of (2.16) belong to the intersection L_0 of the subspaces L_1 and L_2. Hence it follows, in particular, that the right-hand side of (2.16) is a certain linear combination of elements (2.11), i.e. there are numbers $\lambda_1, \ldots, \lambda_k$ such that

$$-\gamma_1\boldsymbol{f}_1-\ldots-\gamma_m\boldsymbol{f}_m=\lambda_1\boldsymbol{e}_1+\ldots+\lambda_k\boldsymbol{e}_k. \tag{2.17}$$

By virtue of linear independence of base elements (2.13), equality (2.17) is only possible when all the coefficients $\gamma_1, \ldots, \gamma_m, \lambda_1, \ldots, \lambda_k$ are equal to zero. But in this case we find from (2.15) that

$$\alpha_1\boldsymbol{g}_1+\ldots+\alpha_l\boldsymbol{g}_l+\beta_1\boldsymbol{e}_1+\ldots+\beta_k\boldsymbol{e}_k=0. \tag{2.18}$$

Because of the linear independence of the base vectors (2.12), equality (2.18) is only possible when all the coefficients $\alpha_1, \ldots, \alpha_l$, $\beta_1, \ldots, \beta_k$ are equal to zero. Thus we have established that equality (2.15) is possible only when all the coefficients $\alpha_1, \ldots, \alpha_l, \beta_1, \ldots, \beta_k$,

* Since in this case the dimension of $\widehat{L}$, equal to $i+k+m$, together with the dimension of L_0, equal to k, is equal to the sum of the dimensions $k+l$ and $k+m$ of the subspaces L_1 and L_2.

$\gamma_1, \ldots, \gamma_m$ are equal to zero, and this proves the linear independence of elements (2.14).

It remains to prove that any element $\boldsymbol{x}$ of the sum $\widehat{L}$ is a certain linear combination of elements (2.14), but this follows immediately from the fact that the element $\boldsymbol{x}$ is (by the definition of $\widehat{L}$) the sum of a certain element $\boldsymbol{x}_1$ of the subspace L_1, which is a linear combination of elements (2.12), and a certain element $\boldsymbol{x}_2$ of the subspace L_2, which is a linear combination of elements (2.13). We have proved the theorem.

Returning to the example considered prior to the formulation of Theorem 2.9, we note that the dimension of L_1 and that of L_2 is two, the dimension of their sum is three, and that of their intersection is unity.

2.3.4. Expansion of a linear space into a direct sum of subspaces. Suppose R_1 and R_2 are two subspaces of the linear n-dimensional space R.

Definition. *We say that the space R is a* ***direct sum*** *of the subspaces R_1 and R_2 if every element $\boldsymbol{x}$ of the space R can be* ***uniquely*** *represented as the sum*

$$\boldsymbol{x}=\boldsymbol{x}_1+\boldsymbol{x}_2 \tag{2.19}$$

of the element $\boldsymbol{x}_1$ of the subspace R_1 and the element $\boldsymbol{x}_2$ of the subspace R_2.

The fact that R is a direct sum of R_1 and R_2 is symbolized as $R=R_1 \oplus R_2$.

The latter equality is usually called the **expansion of the space R into a direct sum of the subspaces R_1 and R_2.**

Thus, the space R of all free vectors (in a three-dimensional space) can be expanded into a direct sum of the subspace R_1 of all vectors parallel to the plane Oxy and the subspace R_2 of all vectors parallel to the axis Oz.

Theorem 2.10. *For the n-dimensional space R to be a direct sum of the subspaces R_1 and R_2 it is sufficient that the intersection of R_1 and R_2 contain only a zero element and that the dimension of R be equal to the sum of the dimensions of the subspaces R_1 and R_2.*

Proof. Let us choose a basis $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_k$ in the subspace R_1 and a basis $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_l$ in the subspace R_2. We shall prove that the union of these bases

$$\boldsymbol{e}_1, \ldots, \boldsymbol{e}_k, \quad \boldsymbol{g}_1, \ldots, \boldsymbol{g}_l \tag{2.20}$$

is the basis of the whole space R. Since by the hypothesis the dimension n of the whole space R is equal to the sum $k+l$ of the dimensions of R_1 and R_2, it is sufficient (by virtue of Theorem 2.5) to prove the linear independence of elements (2.20).

Let us assume that a certain linear combination of elements (2.20) is a zero element, i.e. the equality

$$\alpha_1\boldsymbol{e}_1+\ldots+\alpha_k\boldsymbol{e}_k+\beta_1\boldsymbol{g}_1+\ldots+\beta_l\boldsymbol{g}_l=\boldsymbol{0} \tag{2.21}$$

or

$$\alpha_1 \boldsymbol{e}_1 + \ldots + \alpha_k \boldsymbol{e}_k = -\beta_1 \boldsymbol{g}_1 - \ldots - \beta_l \boldsymbol{g}_l \tag{2.22}$$

holds true. Since the left-hand side of (2.22) is an element of R_1 and the right-hand side is an element of R_2, and the intersection of R_1 and R_2 contains only a zero element, then the left-hand and the right-hand side of (2.25) are a zero element, and this (because of linear independence of the elements of each of the bases $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_k$ and $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_l$) is possible only under the condition

$$\alpha_1 = \ldots = \alpha_k = 0, \quad \beta_1 = \ldots = \beta_l = 0. \tag{2.23}$$

Thus we have found that equality (2.21) is possible only under the condition (2.23), and this proves the linear independence of elements (2.20) and the fact that elements (2.20) form a basis of the whole space R.

Suppose now that $\boldsymbol{x}$ is any element of R. Let us resolve it into components (2.20). We have $\boldsymbol{x} = \lambda_1 \boldsymbol{e}_1 + \ldots + \lambda_k \boldsymbol{e}_k + \mu_1 \boldsymbol{g}_1 + \ldots + \mu_l \boldsymbol{g}_l$ or $\boldsymbol{x} = \boldsymbol{x}_1 + \boldsymbol{x}_2$, where $\boldsymbol{x}_1 = \lambda_1 \boldsymbol{e}_1 + \ldots + \lambda_k \boldsymbol{e}_k$ is an element of R_1 and $\boldsymbol{x}_2 = \mu_1 \boldsymbol{g}_1 + \ldots + \mu_l \boldsymbol{g}_l$ is an element of R_2.

It remains to prove that representation (2.19) is unique. Let us assume that besides (2.19) there holds another representation

$$\boldsymbol{x} = \boldsymbol{x}_1' + \boldsymbol{x}_2', \tag{2.24}$$

where $\boldsymbol{x}_1'$ is an element of R_1 and $\boldsymbol{x}_2'$ is an element of R_2. Subtracting (2.24) from (2.19), we find that $\boldsymbol{0} = \boldsymbol{x}_1 - \boldsymbol{x}_1' + \boldsymbol{x}_2 - \boldsymbol{x}_2'$, or $\boldsymbol{x}_1 - \boldsymbol{x}'_1 = \boldsymbol{x}_2 - \boldsymbol{x}_2'$. Since there is an element of R_1 on the left-hand side of the last equation and an element of R_2 on the right-hand side and since the intersection of R_1 and R_2 contains only a zero element, it follows from that equation that $\boldsymbol{x}_1 - \boldsymbol{x}_1' = \boldsymbol{0}$, $\boldsymbol{x}_2' - \boldsymbol{x}_2 = \boldsymbol{0}$, i.e. $\boldsymbol{x}_1' = \boldsymbol{x}_1$, $\boldsymbol{x}_2 = \boldsymbol{x}_2'$. And this is what we had to prove.

Remark. When the space R is not a straight line but an ordinary sum of the subspaces R_1 and R_2, representation (2.19) of any element $\boldsymbol{x}$ of the space R is also valid but is not, in a general case, unique.

Suppose, for instance, that R is a three-dimensional space of all free vectors, R_1 is a subspace of all vectors which are parallel to the plane Oxy and R_2 is a subspace of all vectors which are parallel to the plane Oxz. We have found in 2.3.3 that R is the sum (but not a direct sum, of course) of the subspaces R_1 and R_2. Let us designate the base vectors parallel to the Ox, Oy and Oz axes as $\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$, respectively, and resolve the arbitrary element $\boldsymbol{x}$ of the space R into the components $\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$. There are real numbers α, β and γ such that $\boldsymbol{x} = \alpha \boldsymbol{i} + \beta \boldsymbol{j} + \gamma \boldsymbol{k}$ so that on one hand $\boldsymbol{x} = \boldsymbol{x}_1 + \boldsymbol{x}_2$, where $\boldsymbol{x}_1 = \alpha \boldsymbol{i} + \beta \boldsymbol{j}$ is an element of R_1 and $\boldsymbol{x}_2 = \gamma \boldsymbol{k}$ is an element of R_2, and on the other hand $\boldsymbol{x} = \boldsymbol{x}_1' + \boldsymbol{x}_2'$, where $\boldsymbol{x}_1' = \beta \boldsymbol{j}$ is an element of R_1 and $\boldsymbol{x}_2' = \alpha \boldsymbol{i} + \gamma \boldsymbol{k}$ is an element of R_2.

2.4. Transformation of Coordinates in Transformation of the Basis of an n-Dimensional Linear Space

2.4.1 Direct and inverse transformations of bases. Suppose $e_1, e_2, \ldots, e_n$ and $e'_1, e'_2, \ldots, e'_n$ are two arbitrary bases of an n-dimensional space R. As any element of the space R, every element $e'_1, e'_2, \ldots, e'_n$ can be resolved into the components $e_1, e_2, \ldots, e_n$. Assume that the elements $e'_1, e'_2 \ldots, e'_n$ can be expressed in terms of $e_1, e_2, \ldots, e_n$ by the formulas

$$\left.\begin{aligned} e'_1 &= a_{11}e_1 + a_{12}e_2 + \ldots + a_{1n}e_n \\ e'_2 &= a_{21}e_1 + a_{22}e_2 + \ldots + a_{2n}e_n, \\ &\ldots\ldots\ldots\ldots \\ e'_n &= a_{n1}e_1 + a_{n2}e_2 + \ldots + a_{nn}e_n. \end{aligned}\right\} \tag{2.25}$$

This means that transition from the first basis $e_1, e_2, \ldots, e_n$ to the second basis $e'_1, e'_2, \ldots, e'_n$ is given by the matrix

$$A = \begin{Vmatrix} a_{11} & a_{12} & & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ \cdot & \cdot & & \cdot \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{Vmatrix}. \tag{2.26}$$

It should be emphasized that the determinant Δ of matrix (2.26) is obviously nonzero* since otherwise, by virtue of theorem 1.7, the rows of that matrix (and, hence, the base elements $e'_1, e'_2, \ldots, e'_n$) would be linearly dependent.

Let us verify that *an inverse change from the second basis $e'_1, e'_2, \ldots, e'_n$ to the first basis $e_1, e_2, \ldots, e_n$ is carried out with the aid of the matrix which is an inverse of the matrix A.*

Recall that the matrix A which is an inverse of the matrix A was introduced in 1.2.7 and has the form

$$B = \begin{Vmatrix} \frac{A_{11}}{\Delta} & \frac{A_{21}}{\Delta} & & \frac{A_{n1}}{\Delta} \\ \frac{A_{12}}{\Delta} & \frac{A_{22}}{\Delta} & & \frac{A_{n2}}{\Delta} \\ \cdot & \cdot & & \cdot \\ \frac{A_{1n}}{\Delta} & \frac{A_{2n}}{\Delta} & & \frac{A_{nn}}{\Delta}, \end{Vmatrix}. \tag{2.27}$$

where Δ is the determinant of the matrix A and A_{ik} is a cofactor of the element a_{ik} of that determinant.

Let us multiply equation (2.25) by the respective cofactors A_{1i}, $A_{2j}, \ldots, A_{nj}$ of the elements of the jth column of the determinant Δ and then add the equations together. As a result, we get (for any

* In 1.2.7. we agreed to call a matrix of this kind non-singular.

number j equal to 1, 2, ..., n)

$$e'_1A_{1j}+e'_2A_{2j}+\ldots+e'_nA_{nj}=\sum_{i=1}^{n} e_i(a_iA_{1j}+a_{2i}A_{2j}+\ldots+a_{ni}A_{nj}).$$

Taking into account that the sum of the products of the elements of the ith column by the respective cofactors of the elements of the jth column is equal to zero for $i\neq j$ and to the determinant Δ for $i=j$*, we get from the last equation

$$e'_1A_{1j}+e'_2A_{2j}+\ldots+e'_nA_{nj}=e_j\Delta,$$

whence $e_j=\frac{A_{1j}}{\Delta}e'_1+\frac{A_{2j}}{\Delta}e'_2+\ldots+\frac{A_{nj}}{\Delta}e'_n$ $(j=1, 2, \ldots, n)$ or, in more detail,

$$\left.\begin{aligned} e_1&=\frac{A_{11}}{\Delta}e'_1+\frac{A_{21}}{\Delta}e'_2+\ldots+\frac{A_{n1}}{\Delta}e'_n,\\ e_2&=\frac{A_{12}}{\Delta}e'_1+\frac{A_{22}}{\Delta}e'_2+\ldots+\frac{A_{n2}}{\Delta}e'_n,\\ &\ldots\ldots\ldots\ldots\ldots\\ e_n&=\frac{A_{1n}}{\Delta}e'_1+\frac{A_{2n}}{\Delta}e'_2+\ldots+\frac{A_{nn}}{\Delta}e'_n.\end{aligned}\right\} \tag{2.28}$$

Formulas (2.28) precisely establish that an inverse transition from the basis $e'_1, e'_2, \ldots, e'_n$ to the basis $e_1, e_2, \ldots, e_n$ can be carried out by the aid of matrix (2.27) which is an inverse of the matrix A. The notation of this inverse will be abbreviated as A^{-1}.

2.4.2. The relation between a change of a basis and a transformation of the corresponding coordinates. Suppose, as before, that the basis $e_1, e_2, \ldots, e_n$ is transformed into the basis $e'_1, e'_2, \ldots, e'_n$ by means of a non-singular matrix (2.26) so that an inverse transformation of the bases is defined by matrix (2.27). Assume, furthermore, that x is an arbitrary element of the linear space R being considered, and $(x_1, x_2 \ldots, x_n)$ are its coordinates with respect to the first basis $e_1, e_2, \ldots, e_n$ and $(x'_1, x'_2, \ldots, x'_n)$ are its coordinates with respect to the second basis $e'_1, e'_2, \ldots, e'_n$, so that

$$x=x'_1e'_1+x'_2e'_2+\ldots+x'_ne'_n=x_1e_1+x_2e_2+\ldots+x_ne_n.$$

Replacing the elements $e_1, e_2, \ldots, e_n$ in this equation by their expressions defined by formulas (2.28), we obtain

$$\begin{aligned} x&=x'_1e'_1+x'_2e'_2+\ldots+x'_ne'_n\\ &=x_1\left(\frac{A_{11}}{\Delta}e'_1+\frac{A_{21}}{\Delta}e'_2+\ldots+\frac{A_{n1}}{\Delta}e'_n\right)\\ &+x_2\left(\frac{A_{12}}{\Delta}e'_1+\frac{A_{22}}{\Delta}e'_2+\ldots+\frac{A_{n2}}{\Delta}e'_n\right)\\ &\ldots\ldots\ldots\ldots\ldots\\ &+x_n\left(\frac{A_{1n}}{\Delta}e'_1+\frac{A_{2n}}{\Delta}e'_2+\ldots+\frac{A_{nn}}{\Delta}e'_n\right).\end{aligned}$$

* See property 4° in 1.2.4.

By virtue of the uniqueness of the resolution into the components $e'_1, e'_2, \ldots, e'_n$, the last equality immediately yields formulas for transition from the coordinates $(x_1, x_2, \ldots, x_n)$ with respect to the first basis to the coordinates $(x'_1, x'_2, \ldots, x'_n)$ with respect to the second basis:

$$\left.\begin{aligned} x'_1 &= \frac{A_{11}}{\Delta} x_1 + \frac{A_{12}}{\Delta} x_2 + \ldots + \frac{A_{1n}}{\Delta} x_n, \\ x'_2 &= \frac{A_{21}}{\Delta} x_1 + \frac{A_{22}}{\Delta} x_2 + \ldots + \frac{A_{2n}}{\Delta} x_n, \\ &\ldots\ldots\ldots\ldots \\ x'_n &= \frac{A_{n1}}{\Delta} x_1 + \frac{A_{n2}}{\Delta} x_2 + \ldots + \frac{A_{nn}}{\Delta} x_n. \end{aligned}\right\} \tag{2.29}$$

Formulas (2.29) show that the transition from the coordinates $(x_1, x_2, \ldots, x_n)$ to the coordinates $(x'_1, x'_2, \ldots, x'_n)$ is carried out by means of the matrix

$$C = \begin{Vmatrix} \frac{A_{11}}{\Delta} & \frac{A_{12}}{\Delta} & & \frac{A_{1n}}{\Delta} \\ \frac{A_{21}}{\Delta} & \frac{A_{22}}{\Delta} & & \frac{A_{2n}}{\Delta} \\ \ldots & \ldots & & \ldots \\ \frac{A_{n1}}{\Delta} & \frac{A_{n2}}{\Delta} & \ldots & \frac{A_{nn}}{\Delta} \end{Vmatrix}$$

transposed to the inverse matrix (2.27)

We arrive at the following conclusion: *if a transition from the first basis to the second is carried out by means of the non-singular matrix A, then the transition from the coordinates of an arbitrary element with respect to the first basis to the coordinates of that element with respect to the second basis is carried out by means of the matrix* $(A^{-1})'$ *transposed to the inverse matrix* A^{-1}.

Chapter 3

SYSTEMS OF LINEAR EQUATIONS

From the elementary course of linear algebra and from analytic geometry the reader is sure to be familiar with a system of two linear equations in two unknowns and with systems of two or three linear equations in three unknowns*. In this chapter we study systems of arbitrary number m of linear equations in the arbitrary number n of unknowns.

We first establish the necessary and sufficient condition for the existence of at least one solution (or, as they say, the **consistency**) of such a system and then seek the whole collection of its solutions.

In 4.4. we consider the case of an approximate specification of all the coefficients of the system and its constant terms which is important for applications. For that case, we present **Tikhonov's regularization method** which makes it possible to find the so-called **normal** (i.e. the closest to the origin) solution of the indicated system with an accuracy corresponding to the accuracy of specification of coefficients and constant terms.

Chapter 6 gives an idea of the numerical (iteration) methods of solving systems of linear equations.

3.1. The Condition for the Consistency of a Linear System

3.1.1. The concept of a system of linear equations and its solutions. In a general case a system of m linear equations in n unknowns (or briefly, a **linear system**) has the form

$$\left.\begin{array}{l} a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n=b_1, \\ a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n=b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m1}x_1+a_{m2}x_2+\ldots+a_{mn}x_n=b_m. \end{array}\right\} \tag{3.1}$$

The designations $x_1, x_2, \ldots, x_n$, are used for the unknowns to be defined (their number n is not assumed to be obligatory equal to the

* See *Analytic Geometry*, Ilyin and Poznyak, Supplement to Chapter 1, Mir Publishers, Moscow, 1984.

number of equations m), the quantities $a_{11}, a_{12}, \ldots, a_{mn}$ called the **coefficients of the system** and the quantities $b_1, b_2, \ldots, b_m$, called the **free** or **constant terms,** are assumed to be known. Every coefficient a_{ij} of the system has two indices, the first of which i indicates the number of the equation and the second j indicates the number of the unknown in which this coefficient is.

System (3.1) is called **homogeneous** if all its constant terms $b_1, b_2, \ldots, b_m$ are equal to zero.

If at least one of the constant terms $b_1, b_2, \ldots, b_m$ is different from zero, then system (3.1) is called **nonhomogeneous.**

System (3.1) is said to be **quadratic** if the number m of its equations is equal to the number n of its unknowns.

The **solution** of system (3.1) is such a collection n of the numbers $c_1, c_2, \ldots, c_n$, which, being substituted into system (3.1) for the unknowns $x_1, x_2, \ldots, x_n$, turns all the equations of the system into identities.

Not every system of the form (3.1) has solutions. Thus the system of linear equations

$$\left.\begin{aligned} x_1+x_2&=1, \\ x_1+x_2&=2 \end{aligned}\right\}$$

obviously has no solutions (since if a solution of the system existed, then, upon its substitution into the system, the left-hand sides of the two equations would contain identical numbers and we would have $1=2$).

The system of equations of the form (3.1) is called **consistent** if it has at least one solution and **inconsistent** if it has no solutions. The consistent system of the form (3.1) can have either one solution or several solutions.

Two solutions $c_1^{(1)}, c_2^{(1)}, \ldots, c_n^{(1)}$ and $c_1^{(2)}, c_2^{(2)} \ldots, c_n^{(2)}$ of the consistent system of the form (3.1) are said to be distinct if at least one of the equalities $c_1^{(1)}=c_1^{(2)}, c_2^{(1)}=c_2^{(2)}, \ldots, c_n^{(1)}=c_n^{(2)}$ is violated.

The consistent system of the form (3.1) is called **determinate** if it has a unique solution.

The consistent system of the form (3.1) is called **indeterminate** if it has at least two distinct solutions.

It is convenient to write the linear system (3.1) in matrix form. For that purpose we use the notion of the product of two matrices (such that the number of the columns of the first matrix is equal to the number of the rows of the second matrix) introduced in 1.1.2. We take the following two matrices as the matrices being multiplied: the matrix

$$A=\begin{Vmatrix} a_{11} & a_{12} & & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ & & \cdot & \\ a_{m1} & a_{m2} & & a_{mn} \end{Vmatrix}, \tag{3.2}$$

containing m rows and n columns and formed from the coefficients in the unknowns (in what follows we call such a matrix a **coefficient matrix** of system (3.1)), and the matrix X consisting of n rows and 1 column, i.e. a column of the form

$$X=\begin{Vmatrix} x_1 \\ x_2 \\ \cdot \\ x_n \end{Vmatrix} \tag{3.3}$$

According to the rule of multiplication of two matrices*, the product AX of matrix (3.2) by matrix (3.3) is a matrix consisting of m rows and 1 column, i.e. a column of the form

$$\begin{Vmatrix} a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n \\ a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n \\ \cdot \quad \cdot \quad \cdot \\ a_{m1}x_1+a_{m2}x_2+\ldots+a_{mn}x_n \end{Vmatrix}. \tag{3.4}$$

The system of equalities (3.1) means that the column (3.4) coincides with the column

$$B=\begin{Vmatrix} b_1 \\ b_2 \\ \cdot \\ b_m \end{Vmatrix}. \tag{3.5}$$

Thus, in matrix notation, system (3.1) can be replaced by one matrix equation

$$AX=B, \tag{3.6}$$

which is equivalent to it and in which the matrices A, X and B are defined by relations (3.2), (3.3) and (3.5). The solution of the matrix equation (3.6) consists in seeking a column (3.3) which turns equation (3.6) into identity when matrix (3.2) and column (3.5) of the right-hand sides are given.

In this and in the following section we shall elucidate the following three questions concerning system (3.1):

(1) the way to find out whether system (3.1) is consistent,

* See 1.1.2, formula (1.4).

(2) the way to find out whether system (3.1) (when it is consistent) is determinate,

(3) the way of seeking the unique solution of the consistent system (3.1) (when it is determinate) and seeking all its solutions (when it is indeterminate).

3.1.2. Non-trivial consistency of a homogeneous system. We begin with considering a homogeneous linear system of the form (3.1), i.e. the system

$$\left.\begin{array}{l} a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n=0, \\ a_{22}x_1+a_{22}x_2+\ldots+a_{2n}x_n=0, \\ \cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot \\ a_{m1}x_1+a_{m2}x_2+\ldots+a_{mn}x_n=0. \end{array}\right\} \tag{3.7}$$

We note that this system is *always consistent* since it always has the so-called **trivial** (or **zero**) solution $x_1=x_2=\ldots=x_n=0$*.

A question arises as to the *conditions under which the homogeneous systeem* (3.7) *has other solutions in addition to the trivial solution* (i.e. *is* ***non-trivial consistent***).

This question is simple to solve. Note that the existence of a non-trivial solution of system (3.7) is equivalent to a linear dependence of the columns of matrix (3.2) (since the linear dependence of the column of matrix (3.2) means that there are numbers $x_1, x_2, \ldots, x_n$, *not all equal to zero*, such that equalities (3.7) hold true).

But by Theorem 1.6 on the base minor, the linear dependence of the columns of matrix (3.2) exists if and only if **not all** the columns of that matrix are basic, i.e. if and only if the order r of the base minor of matrix (3.2) is smaller than the number n of its columns.

We arrive at the following theorem.

Theorem 3.1. *The homogeneous system* (3.7) *has non-trivial solutions if and only if the rank r of matrix* (3.2) *is smaller than the number n of its columns.*

Corollary. *A quadratic homogeneous system*** *has non-trivial solutions if and only if the determinant formed from the coefficients in the unknowns is equal to zero.*

Indeed, in the case of the quadratic homogeneous system (3.7), i.e. for $m=n$, the rank r of matrix (3.2) is smaller than the number $m=n$, if and only if the determinant of that matrix is equal to zero.

---

* Indeed, substituting in system (3.7) zeros for all the unknowns $x_1, x_2, \ldots, x_n$, we turn all the equations of the system into identities.

** I.e., system (3.7) in which the number m of equations is equal to the number n of the unknowns.

3.1.3. The condition of the consistency of a general linear system. Let us now establish the necessary and sufficient condition for the consistency of a general (inhomogeneous in a general case) system of the form (3.1). This system is associated with two matrices: a matrix A defined by relation (3.2) usually called a **system** or **coefficient matrix** of (3.1) (it is formed from the coefficients in the unknowns), and a matrix

$$A_1 = \left\| \begin{matrix} a_{11} & a_{12} & & a_{1n}b_1 \\ a_{21} & a_{22} & & a_{2n}b_2 \\ & \cdot & & \cdot \\ a_{m1} & a_{m2} & \dots & a_{mn}b_m \end{matrix} \right\|, \tag{3.8}$$

which is usually called an **augmented matrix of system** (3.1) (it results from the coefficient matrix by adding column (3.5) of constant terms to that matrix).

The following *fundamental* theorem holds true.

Theorem 3.2 (the Kronecker-Capelli* theorem). *For the linear system* (3.1) *to be consistent, it is necessary and sufficient that the rank of the augmented matrix of that system be equal to that of the coefficient matrix.*

Proof. **(1) Necessity.** Suppose system (3.1) is consistent, i.e. there are numbers $c_1, c_2, \dots, c_n$ such that the equalities

$$\left.\begin{aligned} a_{11}c_1 + a_{12}c_2 + \dots + a_{1n}c_n &= b_1, \\ a_{21}c_1 + a_{22}c_2 + \dots + a_{2n}c_n &= b_2, \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ a_{m1}c_1 + a_{m2}c_2 + \dots + a_{mn}c_n &= b_n \end{aligned}\right\} \tag{3.9}$$

hold true. Let us designate as r the rank of the **matrix of coefficients** (3.1) and consider the linear closure L of the r basic columns of that matrix. By virtue of Theorem 1.6 on the base minor, any column of the matrix of coefficients belongs to the indicated linear closure L. In other words, any column of the augmented matrix (3.8), except for its last column, belongs to the indicated linear closure L.

It follows from equations (3.9) that the last column of the augmented matrix (3.8) also belongs to the linear closure L (since by virtue of equations (3.9), this last column can be linearly expressed in terms of all the columns of the matrix of coefficients and can, therefore, be linearly expressed in terms of its base columns).

Thus, *all the columns of the augmented matrix* (3.8) *belong to the indicated linear closure* L. We have established, in 2.3.2, that the dimension of the indicated linear closure L is equal to r.

---

* Kronecker, Leopold (1823-1891), a German mathematician, Capelli, Alfred (1855-1910), an Italian mathematician.

This means that any $r+1$ columns of the augmented matrix (3.8) are linearly dependent, i.e. the rank of the augmented matrix (equal to the maximum number of linearly independent columns of that matrix) is also equal to the number r. We have proved the necessity.

(2) Sufficiency. Suppose the ranks of the system matrix and the augmented matrix coincide. Then r base columns of the system matrix are also the base columns of the augmented matrix (3.8)*. By Theorem 1.6, on the base minor, the last column of the augmented matrix (3.8) is a certain linear combination of the indicated r base columns. Hence the last column of the augmented matrix (3.8) is also a linear combination of all the columns of the system matrix (3.2)**, i.e. there are numbers $c_1, c_2, \ldots, c_n$ such that equalities (3.9) hold true. The last equalities signify that the numbers $c_1, c_2, \ldots, c_n$ are a solution of system (3.1), i.e. that system is consistent. This completes the proof of the theorem.

3.2. Seeking Solutions of a Linear System

The Kronecker-Capelli theorem establishes the necessary and sufficient condition for a linear system to be consistent but does not present any means of finding the solutions of that system.

We shall occupy ourselves here with seeking the solutions of the linear system (3.1). We shall first consider the simplest case of a quadratic system of linear equations with the determinants of the system matrix different from zero and pass to seeking the collection of the solutions of the general linear system of the form (3.1).

3.2.1. A quadratic system of linear equations with a nonzero determinant of the system matrix. Suppose we are given a quadratic system of linear equations

$$\left.\begin{aligned} a_{11}x_1+a_{12}x_2+\ldots+a_{1n}x_n&=b_1,\\ a_{21}x_1+a_{22}x_2+\ldots+a_{2n}x_n&=b_2,\\ \ldots\ldots\ldots&\ldots\\ a_{n1}x_1+a_{n2}x_2+\ldots+a_{nn}x_n&=b_n \end{aligned}\right\} \quad (3.10)$$

with a nonzero determinant Δ of the system matrix

$$A=\left\|\begin{matrix} a_{11} & a_{12} & & a_{1n}\\ a_{21} & a_{22} & & a_{2n}\\ \cdot & & & \\ a_{n1} & a_{n2} & & a_{nn}\end{matrix}\right\|. \quad (3.11)$$

* Since the indicated r base columns are linearly independent and the number of linearly independent columns of the augmented matrix does not exceed r.

** Without changing the linear combination of r base columns, we can add to it all the nonbase columns with zero factors.

We shall prove that such a system possesses a solution and that the solution is unique. Let us now find that solution. We shall first prove that system (3.10) can have only one solution (i.e. we shall prove the uniqueness of the solution of system (3.10) under the assumption that it exists).

We assume that there are any n numbers $x_1, x_2, \ldots, x_n$ such that being substituted into system (3.10), they turn all the equations of the system into identities (i.e. there is a certain solution $x_1, x_2, \ldots, x_n$ of system (3.10)). Then, multiplying identities (3.10) by the respective cofactors $A_{1j}, A_{2j}, \ldots, A_{nj}$, of the elements of the jth column of the determinant Δ of matrix (3.11) and then adding the resulting identities together, we find (for any number j equal to $1, 2, \ldots, n$)

$$\sum_{i=1}^{n} x_i (a_{1i}A_{1j} + a_{2i}A_{2j} + \ldots + a_{ni}A_{nj}) = b_1A_{1j} + b_2A_{2j} + \ldots + b_nA_{nj}).$$

Taking into account that the sum of the products of the elements of the ith column by the corresponding cofactors of the elements of the jth column is equal to zero at $i \neq j$ and is equal to the determinant Δ of matrix (3.11) at $i = j$*, we get from the last equation

$$x_j\Delta = b_1A_{1j} + b_2A_{2j} + \ldots + b_nA_{nj}. \tag{3.12}$$

We shall use the designation $\Delta_j(b_i)$ *(or, briefly,* Δ_j*) for the determinant resulting from the determinant* Δ *of the coefficient matrix of system* (3.11) *by replacing its jth column by the column of the constant terms* $b_1, b_2, \ldots, b_n$ (leaving all the other columns of Δ intact).

Note that the right-hand side of (3.12) contains the determinant $\Delta_j(b_i)$**, and the equation assumes the form

$$x_j\Delta = \Delta_j \ (j = 1, 2, \ldots, n). \tag{3.13}$$

The determinant Δ of matrix (3.11) being nonzero, equations (3.13) are equivalent to the relations

$$x_j = \frac{\Delta_j}{\Delta} \ (j = 1, 2, \ldots, n). \tag{3.14}$$

We have thus proved that *if the solution* $x_1, x_2, \ldots, x_n$ *of system* (3.10) *with the determinant* Δ *of the system matrix* (3.11) *different from zero exists, then that solution is uniquely defined by formulas* (3.14).

Formulas (3.14) are known as **Cramer's*** rule.**

---

* See property 4° in 1.2.4.

** To ascertain this fact, it is sufficient to write the expansion of the determinant $\Delta_j(b_i)$ according to the elements of the ith column.

*** Cramer, Gabriel (1704-1752), a Swiss mathematician.

It should be emphasized once more that Cramer's formulas have been obtained under the assumption of the existence of a solution and prove its uniqueness.

It remains to prove the existence of a solution of system (3.10). To do that, it is sufficient, by virtue of the Kronecker-Capelli theorem, to prove that the rank of the system matrix (3.11) is equal to that of the augmented matrix*

$$A_1 = \begin{Vmatrix} a_{11} & a_{12} & & a_{1n}b_1 \\ a_{21} & a_{22} & \dots & a_{2n}b_2 \\ \cdot & & & \cdot \quad \cdot \\ a_{n1} & a_{n2} & & a_{nn}b_n \end{Vmatrix}, \tag{3.15}$$

but this is obvious since by the relation $\Delta \neq 0$ the rank of the system matrix is n and the rank of the augmented matrix (3.15) containing n rows cannot exceed the number n and is, therefore, equal to the rank of the system matrix.

We have thus completely proved that *the quadratic system of linear equations* (3.10) *with the determinant of the system matrix different from zero has a solution defined by Cramer's rule* (3.14) *and that the solution is unique.*

It is even easier to verify the assertion we have proved by matrix method. To do that, we replace (as in 1.1) system (3.10) by the matrix equation equivalent to it:

$$AX = B, \tag{3.16}$$

where A is the matrix of the coefficients of (3.11) and X and B are the columns

$$X = \begin{Vmatrix} x_1 \\ x_2 \\ \cdot \\ x_n \end{Vmatrix}, \quad B = \begin{Vmatrix} b_1 \\ b_2 \\ \cdot \quad \cdot \\ b_n \end{Vmatrix},$$

the first of which must be found and the second is given.

Since the determinant Δ of the matrix A is nonzero, there is an inverse matrix A^{-1} (see 1.2.7).

Let us assume that there is a solution of system (3.10), i.e. there is a column X turning the matrix equation (3.16) into identity. Multiplying the indicated identity from the left by the inverse matrix A^{-1}, we get

$$A^{-1}(AX) = A^{-1}B. \tag{3.17}$$

* There is another way of proving the existence of solution of system (3.10) consisting in the verification of the fact that the numbers $x_1, x_2, \dots, x_n$, defined by Cramer's rule (3.14), turn all the equations of system (3.10) into identities.

We must now take into account that by virtue of the associative property of the product of three matrices (see 1.1.2) and according to the relation $A^{-1}A=I$, where I is an identity matrix (see 1.2.7), $A^{-1}(AX)=(A^{-1}A)X=IX=X$, so that we find from (3.17) that

$$X=A^{-1}B. \tag{3.18}$$

Writing out equation (3.18) and taking into account the form of the inverse matrix*, we obtain Cramer's formulas for the elements of the column X.

Thus we have proved that if a solution of the matrix equation (3.16) exists, then it is uniquely defined by relation (3.18) equivalent to Cramer's formulas.

It is easy to verify that the column X defined by relation (3.18) is, indeed, a solution of matrix equation (3.16), i.e. being substituted in that equation, it turns it into identity. In fact, if the column X is defined by equation (3.18), then $AX=A(A^{-1}B)=(AA^{-1})B=IB=B$.

Thus, *if the determinant Δ of the matrix A is nonzero* (i.e. if the matrix is non-singular), *then there is a solution of matrix equation* (3.16), *defined by relation* (3.18) *which is equivalent to Cramer's formulas, and that solution is unique.*

Example. Let us find the solution of the quadratic system of linear equations

$$\left.\begin{aligned} x_1+2x_2+3x_3+4x_4&=30,\\ -x_1+2x_2-3x_3+4x_4&=10,\\ x_2-x_3+x_4&=3,\\ x_1+x_2+x_3+x_4&=10 \end{aligned}\right\}$$

with a nonzero determinant of the system matrix

$$\Delta=\begin{vmatrix} 1 & 2 & 3 & 4\\ -1 & 2 & -3 & 4\\ 0 & 1 & -1 & 1\\ 1 & 1 & 1 & 1 \end{vmatrix}=-4.$$

Since

$$\Delta_1=\begin{vmatrix} 30 & 2 & 3 & 4\\ 10 & 2 & -3 & 4\\ 3 & 1 & -1 & 1\\ 10 & 1 & 1 & 1 \end{vmatrix}=-4,\quad \Delta_2=\begin{vmatrix} 1 & 30 & 3 & 4\\ -1 & 10 & -3 & 4\\ 0 & 3 & -1 & 1\\ 1 & 10 & 1 & 1 \end{vmatrix}=-8,$$

* See formula (1.41) in 1.2.7.

$$\Delta_3=\begin{vmatrix} 1 & 2 & 30 & 4 \\ -1 & 2 & 10 & 4 \\ 0 & 1 & 3 & 1 \\ 1 & 1 & 10 & 1 \end{vmatrix}=-12, \quad \Delta_4=\begin{vmatrix} 1 & 2 & 3 & 30 \\ -1 & 2 & -3 & 10 \\ 0 & 1 & -1 & 3 \\ 1 & 1 & 1 & 10 \end{vmatrix}=-16,$$

the only solution of the system in question has the form $x_1=1$, $x_2=2$, $x_3=3$, $x_4=4$ by virtue of Cramer's rule.

The main significance of Cramer's rule consists in the fact that it gives an explicit expression for the solution of a quadratic system of linear equations (with a nonzero determinant) in terms of the coefficients of the equations and constant terms. The practical application of Cramer's rule is connected with cumbersome calculations ($(n+1)$ determinants of order n must be calculated to solve a system of n equations in n unknowns). It should be added that if the coefficients of the equations and the constant terms are only approximate values of some physical quantities being measured or are rounded off in the process of calculation, then the use of Cramer's rule may lead to major errors and is impractical in a number of cases.

Tikhonov's regularization method, which will be presented in 4.4, makes it possible to find solutions of a linear system with an accuracy corresponding to the accuracy of the specification of the matrix of coefficients of the equations and of the column of constant terms, and in Chapter 6 we shall give an idea of the so-called iteration methods of solving linear systems which make it possible to find solutions for systems of that kind by successive approximations of the unknowns.

We note in conclusion that here we excluded from the consideration the case when the determinant Δ of the system matrix (3.10) turns into zero. This case will be treated in the general theory of systems of m linear equations in n unknowns given in 3.2.2.

3.2.2. Seeking all the solutions of a general linear system. Let us now consider a general system of m linear equations in n unknowns (3.1). We assume that this system is consistent and that the rank of its system and augmented matrices is equal to r. Without losing generality, we can assume that the base minor of system matrix (3.2) is in the top left-hand corner of that matrix (the general case can be reduced to this case by transposing the equations and the unknowns in system (3.1)).

Then the first r rows of both system matrix (3.2) and augmented matrix (3.8) are the base rows of those matrices*, and, according to Theorem 1.6 on the base minor, each of the rows of the augmented

* Since the ranks of the system matrix and the augmented matrix are both equal to r, the base minor of the system matrix is, at the same time, the base minor of the augmented matrix.

matrix (3.8), beginning with the $(r+1)$th row, is a linear combination of the first r rows of that matrix.

In terms of system (3.1), this means that each of the equations of this system, beginning with the $(r+1)$th equation, is a linear combination (i.e. the consequence) of the first r equations of that system (i.e. *every solution of the first r equations of system* (3.1) *turns into identities all the other equations of the system*).

Thus, it is sufficient to find all the solutions of the first r equations of system (3.1). Let us consider the first r equations of system (3.1) writing them in the form

$$\left.\begin{array}{l} a_{11}x_1+a_{12}x_2+\ldots+a_{1r}x_r=b_1-a_{1(r+1)}x_{r+1}-\ldots-a_{1n}x_n, \\ a_{21}x_1+a_{22}x_2+\ldots+a_{2r}x_r=b_2-a_{2(r+1)}x_{r+1}-\ldots-a_{2n}x_n, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{r1}x_1+a_{r2}x_2+\ldots+a_{rr}x_r=b_r-a_{r(r+1)}x_{r+1}-\ldots-a_{rn}x_n. \end{array}\right\} \quad (3.19)$$

If we arbitrarily evaluate the unknowns $x_{r+1}, \ldots, x_n$ as $c_{r+1}, \ldots, c_n$, then system (3.19) turns into a quadratic system of r linear equations in r unknowns $x_1, x_2, \ldots, x_r$, the nonzero base minor of matrix (3.2) being the determinant of the matrix of coefficients of (3.2). By virtue of the results obtained in 3.2.1 system (3.19) has a unique solution defined by Cramer's rule, i.e. for the arbitrarily chosen $c_{r+1}, \ldots, c_n$ there is a unique collection of r numbers $c_1, c_2, \ldots, c_r$ turning all the equations of system (3.19) into identities and defined by Cramer's rule.

To put down this unique solution, we shall agree to use the designation $M_j(d_i)$ for the determinant obtained from the base minor M of matrix (3.2) by replacing its jth column by the column of numbers $d_1, d_2, \ldots, d_i, \ldots, d_r$ (leaving all the other columns of M unchanged). Then, writing the solution of system (3.19) with the aid of Cramer's formulas and using the linear property of the determinant, we get

$$\begin{aligned} c_j &= \frac{1}{M}M_j\,(b_i-a_{i(r+1)}c_{r+1}-\ldots-a_{in}c_n) \\ &= \frac{1}{M}\left[M_j(b_i)-c_{r+1}M_j\,(a_{i(r+1)}-\ldots-c_nM_j\,(a_{in})\right] \qquad (3.20) \end{aligned}$$

$$(j=1,\ 2,\ \ldots,\ r).$$

Formulas (3.20) express the values of the unknowns $x_j=c_j$ $(j=1, 2, \ldots, r)$ in terms of the coefficients in the unknowns, the constant terms and arbitrary parameters $c_{r+1}, \ldots, c_n$.

Let us prove that *formulas* (3.20) *contain any solution of system* (3.1). Indeed, suppose $c_1^0, c_2^0, \ldots, c_r^0, c_{r+1}^0, \ldots, c_n^0$ is an arbitrary solution of the indicated system. Then it is also a solution of system (3.19). But the quantities $c_1^0, c_2^0, \ldots, c_r^0$ appearing in system (3.19) can be uniquely expressed in terms of the quantities

$c_{r+1}^0, \ldots, c_n^0$ by Cramer's formulas (3.20). Thus, for $c_{r+1}=c_{r+1}^0, \ldots$ $\ldots, c_n=c_n^0$ formulas (3.20) give the solution $c_1^0, c_2^0, \ldots, c_r^0, c_{r+1}^0, \ldots, c_n^0$ being considered.

Remark. If the rank r of the system matrix and the augmented matrix of system (3.1) is equal to the number n of the unknowns, then in this case as well relations (3.20) pass into the formulas

$$c_j=\frac{M_j(b_i)}{M} \quad (j=1, 2, \ldots, n).$$

defining the **unique** solution of system (3.1). Thus, *system* (3.1.) *possesses a unique solution* (*i.e. it is determinate*) *under the condition that the rank r of the system matrix and the augmented matrix is equal to the number of the unknowns n* (and is less than the number m of equations or is equal to it).

Example. Let us find all the solutions of the linear system

$$\left.\begin{aligned} x_1-x_2+x_3-x_4&=4,\\ x_1+x_2+2x_3+3x_4&=8,\\ 2x_1+4x_2+5x_3+10x_4&=20,\\ 2x_1-4x_2+x_3-6x_4&=4. \end{aligned}\right\} \tag{3.21}$$

It is easy to ascertain that the rank of both the coefficient matrix and the augmented matrix of this system is equal to 2 (i.e. the system is consistent) and it may be considered that the base minor M is in the top left corner of the system matrix, i.e. $M=\begin{vmatrix} 1 & -1 \\ 1 & 1 \end{vmatrix}=2$. But then, deleting the last two equations and assigning arbitrary values to c_3 and c_4, we get a system

$$\left.\begin{aligned} x_1-x_2&=4-c_3+c_4,\\ x_1+x_2&=8-2c_3-3c_4, \end{aligned}\right\}$$

from which, by virtue of Cramer's rule, we get the values

$$x_1=c_1=6-\frac{3}{2}c_3-c_4, \quad x_2=c_2=2-\frac{1}{2}c_3-2c_4. \tag{3.22}$$

Thus, four numbers

$$\left(6-\frac{3}{2}c_3-c_4,\ 2-\frac{1}{2}c_3-2c_4,\ c_3,\ c_4\right) \tag{3.23}$$

form a solution of system (3.21) for arbitrary values of c_3 and c_4, the row (3.23) containing all the solutions of this system.

3.2.3. The properties of the collection of solutions of a homogeneous system. Let us now discuss a homogeneous system of m

linear equations in n unknowns (3.7) assuming, as before, that matrix (3.2) has a rank equal to r and that the base minor M is in the top left corner of that matrix.

Since in this case all b_i are equal to zero, we get the following formulas instead of (3.20):

$$c_j = -\frac{1}{M}\left[c_{r+1}M_j(a_{i(r+1)}) + \ldots + c_n M_j(a_{in})\right] \quad (j=1;\ 2, \ldots, r) \tag{3.24}$$

which express the values of the unknowns $x_j = c_j\ (j=1, 2, \ldots, r)$ in terms of the coefficients in the unknowns and the arbitrary values $c_{r+1}, \ldots, c_n$. Because of what was proved in 3.2.2, *formulas* (3.24) *contain any solution of homogeneous system* (3.7).

Let us now make sure that the *collection of all solutions of homogeneous system* (3.7) *forms a linear space.*

Suppose $X_1 = (x_1^{(1)}, \ldots, x_n^{(1)})$ and $X_2 = (x_1^{(2)}, \ldots, x_n^{(2)})$ are two arbitrary solutions of homogeneous system (3.7) and λ is any real number. Since every solution of homogeneous system (3.7) is an element of the linear space A^n of all ordered collections of n numbers, it is sufficient to prove that each of the two collections

$$X_1 + X_2 = (x_1^{(1)} + x_1^{(2)}, \ldots, x_n^{(1)} + x_n^{(2)}) \text{ and } \lambda X_1 = (\lambda x_1^{(1)}, \ldots, \lambda x_n^{(1)})$$

is also a solution of homogeneous system (3.7).

Let us consider any equation of (3.7), say, the ith equation, and substitute the elements of the indicated collections for the unknowns in this equation. Taking into account that X_1 and X_2 are solutions of the homogeneous system, we get

$$\sum_{j=1}^{n} a_{ij}[x_j^{(1)} + x_j^{(2)}] = \sum_{j=1}^{n} a_{ij}x_j^{(1)} + \sum_{j=1}^{n} a_{ij}x_j^{(2)} = 0,$$

$$\sum_{j=1}^{n} a_{ij}[\lambda x_j^{(1)}] = \lambda \sum_{j=1}^{n} a_{ij}x_j^{(1)} = 0,$$

and this means that the collections $X_1 + X_2$ and λX_1 are solutions of homogeneous system (3.7).

Thus, the collection of all solutions of homogeneous system (3.7) forms a linear space which we symbolize as R.

Next we find the dimension of that space R and construct a basis in it.

We shall prove that under the assumption that the rank of the matrix of homogeneous system (3.7) is equal to r, *the linear space R of all solutions of homogeneous system* (3.7) *is isomorphic to the linear space A^{n-r} of all the ordered collections of* $(n-r)$ *numbers**.

Let us put every solution $(c_1, \ldots, c_r, c_{r+1}, \ldots, c_n)$ of homogeneous system (3.7) into correspondence with the element

* The space A^m was introduced in Example 3 in 2.1.1.

$(c_{r+1}, \ldots, c_n)$ of the space A^{n-r}. Since the numbers $c_{r+1}, \ldots, c_n$ can be chosen arbitrarily and in each case the solution of system (3.7) is defined uniquely with the aid of formulas (3.24), the correspondence we have established is **one-to-one**. Next we note that if the elements $(c_{r+1}^{(1)}, \ldots, c_n^{(1)})$ and $(c_{r+1}^{(2)}, \ldots, c_n^{(2)})$ of the space A^{n-r} correspond to the elements $(c_1^{(1)}, \ldots, c_r^{(1)}, c_{r+1}^{(1)}, \ldots, c_n^{(1)})$ and $(c_1^{(2)}, \ldots, c_r^{(2)}, c_{r+1}^{(2)}, \ldots, c_n^{(2)})$ of the space R, then it follows immediately from formulas (3.24) that the element $(c_{r+1}^{(1)}+c_{r+1}^{(2)}, \ldots, c_n^{(1)}+c_n^{(2)})$ is associated with the element $(c_1^{(1)}+c_1^{(2)}, \ldots, c_r^{(1)}+c_r^{(2)}, c_{r+1}^{(1)}+c_{r+1}^{(2)}, \ldots, c_n^{(1)}+c_n^{(2)})$, and the element $(\lambda c_{r+1}^{(1)}, \ldots, \lambda c_n^{(1)})$ is associated with the element $(\lambda c_1^{(1)}, \ldots, \lambda c_r^{(1)}, \lambda c_{r+1}^{(1)}, \ldots, \lambda c_n^{(1)})$ for any real λ. We have thus proved that the correspondence we have established is isomorphic.

Thus, *the linear space R of all solutions of homogeneous system* (3.7) *in n unknowns and with the rank of the system matrix equal to r is isomorphic to the space A^{n-r} and, hence, has a dimension $n-r$.*

Any collection of $(n-r)$ linearly independent solutions of homogeneous system (3.7) *form* (by virtue of Theorem 2.5) *a basis in the space R of all solutions and is known as a* ***fundamental collection of solutions*** *of homogeneous system* (3.7).

To construct a fundamental collection of solutions, we can proceed from any basis of the space A^{n-r}. By virtue of isomorphism, the collection of solutions of system (3.7) corresponding to this basis is linearly independent and is, therefore, a fundamental collection of solutions.

We must isolate the fundamental collection of solutions of system (3.7) corresponding to the simplest basis $\boldsymbol{e}_1=(1, 0, 0, \ldots, 0)$, $\boldsymbol{e}_2=(0, 1, 0, \ldots, 0), \ldots, \boldsymbol{e}_{n-r}=(0, 0, 0, \ldots, 1)$ of the space A^{n-r} and known as a **normal fundamental collection of solutions** of homogeneous system (3.7).

Under the assumptions made above concerning the rank and the position of the base minor, according to formulas (3.24), the normal fundamental collection of solutions of homogeneous system (3.7) has the form

$$\left.\begin{aligned}
X_1 &= \left(-\frac{M_1(a_{i\,(r+1)})}{M}, \ldots, -\frac{M_r(a_{i\,(r+1)})}{M}, 1, 0, \ldots, 0\right),\\
X_2 &= \left(-\frac{M_1(a_{i\,(r+2)})}{M}, \ldots, -\frac{M_r(a_{i(r+2)})}{M}, 0, 1, \ldots, 0\right),\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
X_{n-r} &= \left(-\frac{M_1(a_{in})}{M}, \ldots, -\frac{M_r(a_{in})}{M}, 0, 0, \ldots, 1\right).
\end{aligned}\right\} \qquad (3.25)$$

By the definition of a basis, any solution X of homogeneous

system (3.7) can be represented as

$$X = C_1X_1 + C_2X_2 + \ldots + C_{n-r}X_{n-r}, \tag{3.26}$$

where $C_1, C_2, \ldots, C_{n-r}$ are some constants. Since formula (3.26) includes any solution of homogeneous system (3.7), this formula gives a **general solution** of the homogeneous system being considered.

Example. Let us consider a homogeneous system of equations

$$\left.\begin{aligned} x_1 - x_2 + x_3 - x_4 &= 0, \\ x_1 + x_2 + 2x_3 + 3x_4 &= 0, \\ 2x_1 + 4x_2 + 5x_3 + 10x_4 &= 0, \\ 2x_1 - 4x_2 + x_3 - 6x_4 &= 0, \end{aligned}\right\} \tag{3.27}$$

corresponding to nonhomogeneous system (3.21) discussed in the example at the end of 3.2.2. We have found there that the rank r of the matrix of that system is two and have taken the minor positioned at the top left corner of the indicated matrix as a base minor.

Repeating the arguments presented at the end of 3.2.2, we get, instead of formulas (3.22), relations

$$c_1 = -\frac{3}{2}c_3 - c_4, \quad c_2 = -\frac{1}{2}c_3 - 2c_4$$

valid for the arbitrarily chosen c_3 and c_4. With the aid of these relations (setting first $c_3 = 1$, $c_4 = 0$ and then $c_3 = 0$, $c_4 = 1$), we obtain a normal fundamental collection of two solutions of system (3.27):

$$X_1 = \left(-\frac{3}{2}, -\frac{1}{2}, 1, 0\right), \; X_2 = (-1, -2, 0, 1).$$

The general solution of homogeneous system (3.27) has the form

$$X = C_1\left(-\frac{3}{2}, -\frac{1}{2}, 1, 0\right) + C_2(-1, -2, 0, 1), \tag{3.28}$$

where C_1 and C_2 are arbitrary constants.

Now we can establish a connection between the solutions of nonhomogeneous linear system (3.1) and the corresponding homogeneous system (3.7)*. Let us prove the following **two assertions.**

1°. *The sum of any solution of nonhomogeneous system* (3.1) *and any solution of the corresponding homogeneous system* (3.7) *is a solution of system* (3.1).

Indeed, if $c_1, \ldots, c_n$ is a solution of system (3.1) and $d_1, \ldots, d_n$ is a solution of the corresponding homogeneous system (3.7), then substituting the numbers $c_1 + d_1, \ldots, c_n + d_n$ for the unknowns

* With the same coefficients in the unknowns.

in any (say, ith) equation of system (3.1), we get

$$\sum_{j=1}^{n} a_{ij}(c_j+d_j)=\sum_{j=1}^{n} a_{ij}c_j+\sum_{j=1}^{n} a_{ij}d_j=b_i+0=b_i.$$

And that is what we set out to prove.

2°. *The difference between two arbitrary solutions of nonhomogeneous system* (3.1) *is a solution of the corresponding homogeneous system* (3.7).

Indeed, if $c_1', \ldots, c_n'$ and $c_1'', \ldots, c_n''$ are two arbitrary solutions of system (3.1), then, substituting the numbers $c_1'-c_1'', \ldots, c_n'-c_n''$, for the unknowns in any (say, ith) equation of system (3.7), we get

$$\sum_{j=1}^{n} a_{ij}(c_j'-c_j'')=\sum_{j=1}^{n} a_{ij}c_j'-\sum_{j=1}^{n} a_{ij}c_j''=b_i-b_i=0.$$

And that is what we wished to prove.

It follows from the assertion we have proved that *having found one solution of nonhomogeneous system* (3.1) *and adding it to every solution of the corresponding homogeneous system* (3.7), *we can get all the solutions of nonhomogeneous system* (3.1).

In other words, *the sum of a particular solution of the nonhomogeneous system* (3.1) *and the general solution of the corresponding homogeneous system* (3.7) *yields a general solution of nonhomogeneous system* (3.1).

It is natural to take as a particular solution of nonhomogeneous system (3.1) its solution*

$$X_0=\left(\frac{M_1(b_i)}{M}, \ldots, \frac{M_r(b_i)}{M}, 0, 0, \ldots, 0\right), \tag{3.29}$$

which results if we set all the numbers $c_{r+1}, \ldots, c_n$ in formulas (3.20) equal to zero. Adding that particular solution to the general solution (3.26) of the corresponding homogeneous system, we get the following expression for the general solution of the nonhomogeneous system (3.1):

$$X=X_0+C_1X_1+C_2X_2+\ldots+C_{n-r}X_{n-r}. \tag{3.30}$$

In this expression X_0 is a particular solution (3.29), $C_1, C_2, \ldots, C_{n-r}$ are arbitrary constants, and $X_1, X_2, \ldots, X_{n-r}$ are elements of the normal fundamental collection of solutions (3.25) of the corresponding homogeneous system.

Thus, for nonhomogeneous system (3.21) considered at the end of 3.2.2, a particular solution of form (3.29) is $X_0=(6, 2, 0, 0)$. Adding this particular solution to the general solution (3.28)

* We assume here, as before, that the ranks of the coefficient and the augmented matrix of system (3.1) are equal to r and that the base minor is in the top left corner of those matrices.

of the corresponding homogeneous system (3.27), we get the following general solution of nonhomogeneous system (3.21):

$$X=(6,\ 2,\ 0,\ 0)+C_1\left(-\frac{3}{2},\ -\frac{1}{2},\ 1,\ 0\right)+C_2(-1,\ -2,\ 0,\ 1).$$

(Here C_1 and C_2 are arbitrary constants.)

3.2.4. Concluding remarks on solution of linear systems. The methods of solving linear systems discussed in detail in the previous subsections are held up by the necessity of calculating the rank of the matrix and finding its base minor. When the base minor is known, the solution reduces to the technique of calculating the determinants and using Cramer's rule.

The following **rule** can be used to calculate the rank of a matrix: *in calculating the rank of a matrix, we must pass from minors of lower orders to those of higher orders; and, if we have found the nonzero minor M of order k, then we have only to calculate the minors of order $(k+1)$,* ***bordering**** *that minor M; if all the bordering minors of order $(k+1)$ are equal to zero, then the rank of the matrix is equal to k***.

Here is another rule for calculating the rank of a matrix. Note that we can do **three simple operations** on the rows (columns) of a matrix without changing the rank of the matrix: (1) transposition of two rows (or two columns), (2) multiplication of a row (or a column) by any nonzero factor, (3) addition to a row (column) of an arbitrary linear combination of the other rows (columns)***.

We say that the matrix $\| a_{ij} \|$ containing m rows and n columns is of a **diagonal** form if all the elements different from $a_{11}, a_{22}, \ldots, a_{rr}$, where $r=\min\{m, n\}$, are equal to zero. The rank of such a matrix is evidently equal to r.

Let us verify that *by means of three simple operations we can reduce any matrix*

$$A=\begin{Vmatrix} a_{11} & & a_{1n} \\ & & \cdot \\ a_{m1} & & a_{mn} \end{Vmatrix} \tag{3.31}$$

* I.e., containing the minor M in their interior.

** Indeed, in the indicated case all the rows (columns) of the matrix belong to the linear closure of its k rows (columns) on whose intersection the minor M is located, and the dimension of the linear closure is k.

*** These three operations do not change the rank of the matrix because operations (1) and (2) do not change the maximum number of the linearly independent rows (columns) of the matrix, and operation (3) possesses the property according to which the linear closure of all the rows (columns) existing before the operation coincides with the linear closure of all the rows (columns) resulting from the operation carried out.

to a diagonal form (and this enables us to calculate its rank).

Indeed, if all the elements of matrix (3.31) are equal to zero, this means that the matrix has been reduced to a diagonal form. Now if matrix (3.31) has nonzero elements, then by interchanging two rows and two columns we can make the element a_{11} differ from zero. Then, multiplying the first row of the matrix by a_{11}^{-1}, we turn the element a_{11} into unity. Subtracting then the first column multiplied by a_{1j} from the jth column of the matrix (for $j=2, 3, \ldots, n$) and after that subtracting the first row multiplied by a_{i1} from the ith row (for $i=2, 3, \ldots, n$), we get a matrix of the form

$$\left\| \begin{array}{c|ccc} 1 & 0 & & 0 \\ \hline 0 & a'_{22} & & a'_{2n} \\ \cdot & \cdot & & \cdot \\ 0 & a'_{m2} & \ldots & a'_{mn} \end{array} \right\|$$

instead of (3.31). Performing the indicated operations on the matrix in the frame and proceeding as before, we get, after a finite number of steps, a diagonal matrix.

The methods of solving linear systems, using, in the final analysis, the apparatus of Cramer's rule, which we have discussed in the previous subsections, can lead to major errors when the values of the coefficients of the equations and constant terms are approximate or when these values are rounded off in the process of calculation. First and foremost, it concerns the case when a matrix corresponding to the principal determinant (or to the base minor) is **ill-conditioned** (i.e. when "small" changes in the elements of that matrix are associated with "great" changes in the elements of the inverse matrix). Naturally, the solution of a linear system will be **unstable** in this case (i.e. "small" changes in the values of the coefficients of the equations and constant terms will be associated with "great" changes in the solutions). These circumstances necessitate an elaboration of other (distinct from Cramer's rule) theoretical algorithms for seeking solutions and numerical methods of solving linear systems.

In 4.4 we shall get acquainted with Tikhonov's **regularization method** for seeking the so-called **normal** (i.e. the closest to the origin) solution of a linear system.

Chapter 6 contains the fundamental knowledge of the so-called **iteration methods** of solution of linear systems making it possible to solve such systems by means of successive approximations of the unknowns.

Chapter 4

EUCLIDEAN SPACES

The reader is sure to be acquainted, from the course of analytic geometry, with the concept of a scalar product of two free vectors and with the four principal properties of that scalar product. In this chapter we shall study linear spaces of any nature for whose elements in any way (no matter which) a rule is defined associating any two elements with a number called the **scalar product** of those elements. It is only essential that the rule should possess the same four properties as the rule of forming a scalar product of two free vectors. Linear spaces in which the indicated rule is defined are called **Euclidean spaces.** In this chapter, we elucidate the main properties of arbitrary Euclidean spaces.

4.1. A Real Euclidean Space and Its Simplest Properties

4.1.1. Definition of a real Euclidean space. *A real linear space R is called a* ***real Euclidean*** *space (or, simply, a* ***Euclidean space****) if the following two requirements are met:*

I. There is a rule which puts any two elements $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *of that space into correspondence with a real number called the* ***scalar product*** *of those elements and symbolized as* $(\boldsymbol{x}, \boldsymbol{y})$.

II. This rule is subordinated to the following four axioms:

1°. $(\boldsymbol{x}, \boldsymbol{y})=(\boldsymbol{y}, \boldsymbol{x})$ (commutative property or symmetry).

2°. $(\boldsymbol{x}_1+\boldsymbol{x}_2, \boldsymbol{y})=(\boldsymbol{x}_1, \boldsymbol{y})+(\boldsymbol{x}_2, \boldsymbol{y})$ (distributive property).

3°. $(\lambda\boldsymbol{x}, \boldsymbol{y})=\lambda(\boldsymbol{x}, \boldsymbol{y})$ *for any real* λ.

4°. $(\boldsymbol{x}, \boldsymbol{x})>0$ if $\boldsymbol{x}$ *is a nonzero element,* $(\boldsymbol{x}, \boldsymbol{x})=0$ *if* $\boldsymbol{x}$ *is a zero element.*

It should be emphasized that when introducing the notion of a Euclidean space we abstract ourselves not only from the nature of the objects in question but also from the specific features of the rules for the formation of the sum of elements, the product of an element by a scalar and the scalar product of elements (it is only essential that the rules satisfy the eight axioms of a linear space and the four axioms of a scalar product).

Now if the nature of the objects in question and the form of the indicated rules are specified, then the Euclidean space is said to be **specific.**

Here are some examples of specific Euclidean spaces.

Example 1. Let us consider the linear space B_3 of all free vectors. We define the scalar product of any two vectors as we did it in analytic geometry (i.e. as the product of the lengths of these vectors by the cosine of the angle between them). In the course of analytic geometry we proved the validity of axioms 1°-4° for a scalar product thus defined*. Hence, the space B_3 with a scalar product thus defined is a Euclidean space.

Example 2. Let us consider the finite-dimensional linear space $C[a, b]$ of all functions $x(t)$ defined and continuous on the interval $a \leqslant t \leqslant b$. The scalar product of two such functions $x(t)$ and $y(t)$ is defined as an integral (from a to b) of the product of these functions

$$\int_a^b x(t)\, y(t)\, dt. \tag{4.1}$$

The validity of axioms 1°-4° for the scalar product thus defined can be simply verified. Indeed, the validity of axiom 1° is obvious, the validity of axioms 2° and 3° follows from the linear properties of a definite integral; the validity of axiom 4° follows from the fact that the integral $\int_a^b x^2(t)\, dt$ of the continuous nonnegative function $x^2(t)$ is nonnegative and vanishes only if that function is identically zero on the interval $a \leqslant t \leqslant b$** (i.e. is a zero element of the space being considered). Thus, the space $C[a, b]$ with the scalar product thus defined is an infinite-dimensional Euclidean space.

Example 3. This example of a Euclidean space gives an n-dimensional linear space A^n of ordered collections of n real numbers, the scalar product of whose any two elements $\boldsymbol{x}=(x_1, x_2, \ldots, x_n)$ and $\boldsymbol{y}=(y_1, y_2, \ldots, y_n)$ is defined by the equation

$$(\boldsymbol{x}, \boldsymbol{y})=x_1y_1+x_2y_2+\ldots+x_ny_n. \tag{4.2}$$

The validity of axiom 1° for the scalar product thus defined is obvious, the validity of axioms 2° and 3° can be easily verified (it is sufficient to recall the definition of the operations of addition of elements and their multiplication by scalars:

$$(x_1, x_2, \ldots, x_n)+(y_1, y_2, \ldots, y_n)$$
$$=(x_1+y_1, x_2+y_2, \ldots, x_n+y_n),$$
$$\lambda(x_1, x_2, \ldots, x_n)=(\lambda x_1, \lambda x_2, \ldots, \lambda x_n);$$

* See *Analytic Geometry* by Ilyin and Poznyak, Mir Publishers, Moscow, 1984.

** See *Fundamentals of Mathematical Analysis*, Ilyin and Poznyak, Mir Publishers, Moscow, 1982, Part 1, properties 1° and 2° in 10.6.1.

and, finally, the validity of axiom 4° follows from the fact that $(\boldsymbol{x}, \boldsymbol{x})=x_1^2+x_2^2+\ldots+x_n^2$ is always a nonnegative number and vanishes only under the condition $x_1=x_2=\ldots=x_n=0$.

The Euclidean space considered in this example is often symbolized as E^n.

Example 4. In the same linear space A^n we introduce a scalar product of any two elements $\boldsymbol{x}=(x_1, x_2, \ldots, x_n)$ and $\boldsymbol{y}=(y_1, y_2, \ldots, y_n)$ not by relation (4.2) but in some other, more general, way.

For that purpose we consider a square matrix of order n:

$$A=\begin{Vmatrix} a_{11} & a_{12} & & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ \cdot & & & \cdot \\ a_{n1} & a_{n2} & & a_{nn} \end{Vmatrix}. \tag{4.3}$$

Let us form, with the aid of matrix (4.3), a homogeneous polynomial of order 2 with respect to n variables $x_1, x_2, \ldots, x_n$

$$\sum_{i=1}^{n}\sum_{k=1}^{n} a_{ik}x_ix_k. \tag{4.4}$$

Forestalling events, we note that such a polynomial is called a **quadratic form** (generated by matrix (4.3))*.

Quadratic form (4.4) is said to be **positive definite** if it assumes strictly positive values for all the values of the variables $x_1, x_2, \ldots, x_n$ which are simultaneously nonzero**. Since for $x_1=x_2=\ldots=x_n=0$ the quadratic form (4.4) is evidently equal to zero, we can say that a *positive definite quadratic form vanishes only under the condition* $x_1=x_2=\ldots=x_n=0$.

Let us require that matrix (4.3) satisfy the following two conditions:

1°. It should generate a positive definite quadratic form (4.4).

2°. It should be symmetric (about the principal diagonal), i.e. satisfy the condition $a_{ik}=a_{ki}$ for all $i=1, 2, \ldots, n$ and all $k=1, 2, \ldots, n$.

Let us use matrix (4.3) satisfying conditions 1° and 2° to define the scalar product of any two elements $\boldsymbol{x}=(x_1, x_2, \ldots, x_n)$ and $\boldsymbol{y}=(y_1, y_2, \ldots, y_n)$ of the space A^n by the relation

$$(\boldsymbol{x}, \boldsymbol{y})=\sum_{i=1}^{n}\sum_{k=1}^{n} a_{ik}x_iy_k. \tag{4.5}$$

It is easy to verify the validity of axioms 1°-4° for the scalar

* Quadratic forms are systematically studied in Chapter 7 of this book.

** In Chapter 7 of this book we give a necessary and sufficient condition for a quadratic form to be positive definite.

product thus defined. Indeed, Axioms 2° and 3° are, evidently, valid for arbitrary matrix (4.3), the validity of axiom 1° follows from the condition of symmetry of matrix (4.3), and the validity of axiom 4° follows from the fact that the quadratic form (4.4), which is a scalar product $(\boldsymbol{x}, \boldsymbol{x})$, is positive definite.

Thus, the space A^n with a scalar product defined by equation (4.5) is a Euclidean space under the condition of symmetry of matrix (4.3) and positive definiteness of the quadratic form it generates.

If we take an identity matrix as matrix (4.3), then relation (4.4) turns into (4.2) and we get the Euclidean space E^n considered in Example 3.

4.1.2. The simplest properties of an arbitrary Euclidean space. The properties that will be established here are valid for an arbitrary Euclidean space of both finite and infinite dimensionality.

Theorem 4.1. *The inequality*

$$(\boldsymbol{x}, \boldsymbol{y})^2 \leqslant (\boldsymbol{x}, \boldsymbol{x})(\boldsymbol{y}, \boldsymbol{y}), \tag{4.6}$$

known as the ***Cauchy-Bunyakovsky inequality,*** *is valid for any two elements* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *of an arbitrary Euclidean space.*

Proof. The inequality $(\lambda\boldsymbol{x}-\boldsymbol{y}, \lambda\boldsymbol{x}-\boldsymbol{y}) \geqslant 0$ is valid for any real number λ by virtue of axiom 4° of a scalar product. According to axioms 1°-3°, we can rewrite the last inequality as

$$\lambda^2(\boldsymbol{x}, \boldsymbol{x}) - 2\lambda(\boldsymbol{x}, \boldsymbol{y}) + (\boldsymbol{y}, \boldsymbol{y}) \geqslant 0.$$

The necessary and sufficient condition for the last quadratic trinomial to be nonnegative is the nonpositiveness of its discriminant, i.e. the inequality*

$$(\boldsymbol{x}, \boldsymbol{y})^2 - (\boldsymbol{x}, \boldsymbol{x})(\boldsymbol{y}, \boldsymbol{y}) \leqslant 0. \tag{4.7}$$

Inequality (4.6) follows immediately from (4.7). We have proved the theorem.

Our immediate task is to introduce the notion of the **norm** (or **length**) of every element in an arbitrary Euclidean space. For that purpose, we introduce the notion of a normed linear space.

Definition. *The linear space R is said to be* ***normed*** *if the following two requirements are satisfied*:

I. *There is a rule which puts every element* $\boldsymbol{x}$ *of the space R into correspondence with a real number called the* ***norm*** *(or the* ***length****) of the indicated element and is designated as* $\|\boldsymbol{x}\|$.

II. *The indicated rule is subordinated to the following three axioms*:

1°. $\|\boldsymbol{x}\| > 0$ if $\boldsymbol{x}$ *is a nonzero element*; $\|\boldsymbol{x}\| = 0$ *if* $\boldsymbol{x}$ *is a zero element.*

---

* In the case $(\boldsymbol{x}, \boldsymbol{x}) = 0$, the quadratic trinomial degenerates into a linear function, but in that case the element $\boldsymbol{x}$ is equal to zero so that $(\boldsymbol{x}, \boldsymbol{y}) = 0$ and inequality (4.7) is also valid.

2°. $\| \lambda x \| = | \lambda | \, \| x \|$ *for any element* x *and any real number* λ.

3°. *The following inequality is valid for any two elements* x *and* y:

$$\| x+y \| \leqslant \| x \| + \| y \|, \tag{4.8}$$

called ***triangle inequality*** (or ***Minkowski's inequality***).

Theorem 4.2. *Every Euclidean space is normed if the norm of its any element* x *is defined by the equation*

$$\| x \| = \sqrt{(x, x)}. \tag{4.9}$$

Proof. It is sufficient to prove that axioms 1°-3° from the definition of a normed space hold true for the norm defined by relation (4.9).

The validity of axiom 1° for the norm follows immediately from axiom 4° for a scalar product. The validity of axiom 2° for the norm follows almost directly from axioms 1° and 3° for a scalar product.

It remains to verify the validity of axiom 3° , i.e. inequality (4.8), for the norm. We proceed from the Cauchy-Bunyakovsky inequality (4.6), which we rewrite as

$$|(x, y)| \leqslant \sqrt{(x, x)} \sqrt{(y, y)}. \tag{4.7'}$$

With the aid of the last inequality, axioms 1°-4° for a scalar product and the definition of a norm, we obtain

$$\| x+y \| = \sqrt{(x+y, x+y)} = \sqrt{(x, x)+2(x, y)+(y, y)}$$
$$\leqslant \sqrt{(x, x)+\sqrt{(x, x)} \cdot \sqrt{(y, y)}+(y, y)}$$
$$= \sqrt{[\sqrt{(x, x)}+\sqrt{(y, y)}]^2} = \sqrt{(x, x)}+\sqrt{(y, y)} = \| x \| + \| y \|.$$

And this completes the proof of the theorem.

Corollary. *In every Euclidean space with the norm of the elements defined by relation* (4.9), *the triangle inequality* (4.8) *is valid for any two elements* x *and* y.

Note that in any **real** Euclidean space the notion of the angle between two arbitrary elements x and y of the space can be introduced. By complete analogy with vector algebra, we take the angle (varying from 0 to π) whose cosine is defined by the relation

$$\cos \varphi = \frac{(x, y)}{\| x \| \, \| y \|} = \frac{(x, y)}{\sqrt{(x, x)} \sqrt{(y, y)}}$$

as the angle φ between the elements x and y.

The definition of the angle we have given is correct since,

by virtue of the Cauchy-Bunyakovsky inequality (4.7′), the fraction on the right-hand side of the last equation does not exceed unity in absolute value.

Next *we agree to call two arbitrary elements* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *of the Euclidean space* $\boldsymbol{E}$ ***orthogonal*** *if the scalar product of those elements* $(\boldsymbol{x}, \boldsymbol{y})$ *is zero* (in that case the angle φ between the elements $\boldsymbol{x}$ and $\boldsymbol{y}$ is zero).

Appealing again to vector algebra, we call the sum $\boldsymbol{x}+\boldsymbol{y}$ of two **orthogonal elements** $\boldsymbol{x}$ and $\boldsymbol{y}$ the **hypotenuse** of a right triangle constructed on the elements $\boldsymbol{x}$ and $\boldsymbol{y}$.

Note that the **Pythagoras theorem** holds true for every Euclidean space: *the square of the hypotenuse is equal to the sum of the squares of the lengths of the legs.* Indeed, since $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthogonal and $(\boldsymbol{x}, \boldsymbol{y})=0$, we have

$$\|\boldsymbol{x}+\boldsymbol{y}\|^2=(\boldsymbol{x}+\boldsymbol{y}, \boldsymbol{x}+\boldsymbol{y})=(\boldsymbol{x}, \boldsymbol{x})+2(\boldsymbol{x}, \boldsymbol{y})+(\boldsymbol{y}, \boldsymbol{y})$$
$$=(\boldsymbol{x}, \boldsymbol{x})+(\boldsymbol{y}, \boldsymbol{y})=\|\boldsymbol{x}\|^2+\|\boldsymbol{y}\|^2$$

by virtue of the axioms and the definition of a norm.

This result can be generalized to n pairwise orthogonal elements $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$: if $\boldsymbol{z}=\boldsymbol{x}_1+\boldsymbol{x}_2+\ldots+\boldsymbol{x}_n$, then

$$\|\boldsymbol{z}\|^2=(\boldsymbol{x}_1+\boldsymbol{x}_2+\ldots+\boldsymbol{x}_n, \boldsymbol{x}_1+\boldsymbol{x}_2+\ldots+\boldsymbol{x}_n)$$
$$=(\boldsymbol{x}_1, \boldsymbol{x}_1)+(\boldsymbol{x}_2, \boldsymbol{x}_2)+\ldots+(\boldsymbol{x}_n, \boldsymbol{x}_n)$$
$$=\|\boldsymbol{x}_1\|^2+\|\boldsymbol{x}_2\|^2+\ldots+\|\boldsymbol{x}_n\|^2.$$

In conclusion we shall write the norm, the Cauchy-Bunyakovsky inequality and the triangle inequality in each specific Euclidean space discussed in 4.1.1.

In a Euclidean space of all free vectors with the ordinary definition of the scalar product, the norm of the vector $\boldsymbol{a}$ coincides with its length $|\boldsymbol{a}|$, the Cauchy-Bunyakovsky inequality can be reduced to the form $(\boldsymbol{a}, \boldsymbol{b})^2 \leqslant |\boldsymbol{a}|^2 |\boldsymbol{b}|^2$* and the triangle inequality can be reduced to the form $|\boldsymbol{a}+\boldsymbol{b}| \leqslant |\boldsymbol{a}|+|\boldsymbol{b}|$**.

In the Euclidean space $C[a, b]$ of all functions $x=x(t)$ with the scalar product (4.1) continuous on the interval $a \leqslant t \leqslant b$, the norm of the element $x=x(t)$ is $\sqrt{\int_a^b x^2(t)\,dt}$, and the Cauchy-Bunyakovsky

---

* For the scalar product of the vectors $(\boldsymbol{a}, \boldsymbol{b})=|\boldsymbol{a}|\,|\boldsymbol{b}| \cos\varphi$ this inequality trivially follows from the fact that $\cos^2\varphi \leqslant 1$.

** If we add the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ together by the rule of triangle, then this inequality can be trivially reduced to the statement that one side of a triangle does not exceed the sum of its other two sides.

inequality and the triangle inequality have the forms

$$\left[\int_a^b x(t)\,y(t)\,dt\right]^2 \leqslant \int_a^b x^2(t)\,dt \int_a^b y^2(t)\,dt,$$

$$\sqrt{\int_a^b [x(t)+y(t)]^2\,dt} \leqslant \sqrt{\int_a^b x^2(t)\,dt} + \sqrt{\int_a^b y^2(t)\,dt}.$$

These two inequalities play a significant role in various divisions of mathematical analysis.

In the Euclidean space E^n of ordered collections of n real numbers with scalar product (4.2) the norm of any element $\boldsymbol{x}=(x_1, x_2, \ldots, x_n)$ is

$$\|\boldsymbol{x}\| = \sqrt{x_1^2+x_2^2+\ldots+x_n^2},$$

and the Cauchy-Bunyakovsky inequality and the triangle inequality have the forms

$$(x_1y_1+x_2y_2+\ldots+x_ny_n)^2 \leqslant (x_1^2+x_2^2+\ldots+x_n^2)(y_1^2+y_2^2+\ldots+y_n^2),$$

$$\sqrt{(x_1+y_1)^2+(x_2+y_2)^2+\ldots+(x_n+y_n)^2} \leqslant \sqrt{x_1^2+x_2^2+\ldots+x_n^2}+\sqrt{y_1^2+y_2^2+\ldots+y_n^2}.$$

Finally, in a Euclidean space of ordered collections of n real numbers with scalar product (4.5), the norm of any element $\boldsymbol{x}=(x_1, x_2, \ldots, x_n)$ is*

$$\|\boldsymbol{x}\| = \sqrt{\sum_{i=1}^n \sum_{k=1}^n a_{ik}x_ix_k}$$

and the Cauchy-Bunyakovsky and triangle inequalities have the forms

$$\left(\sum_{i=1}^n \sum_{k=1}^n a_{ik}x_iy_k\right)^2 \leqslant \left(\sum_{i=1}^n \sum_{k=1}^n a_{ik}x_ix_k\right)\left(\sum_{i=1}^n \sum_{k=1}^n a_{ik}y_iy_k\right),$$

$$\sqrt{\sum_{i=1}^n \sum_{k=1}^n a_{ik}(x_i+y_i)(x_k+y_k)} \leqslant \sqrt{\sum_{i=1}^n \sum_{k=1}^n a_{ik}x_ix_k} + \sqrt{\sum_{i=1}^n \sum_{k=1}^n a_{ik}y_iy_k}.$$

* Recall that in this case matrix (4.3) is symmetric and generates a positive definite quadratic form (4.4).

4.2. An Orthonormal Basis of a Finite-Dimensional Euclidean Space

We shall study here Euclidean spaces of a **finite** dimension n. The extension of the results discussed here to infinite-dimensional Euclidean spaces does not enter the scope of this book and is the subject for special discussion. (Such spaces are discussed in Chapters 10 and 11 of our *Fundamentals of Mathematical Analysis*, Part 2.)

4.2.1. The concept of an orthonormal basis and its existence. In Chapter 2 we introduced the notion of a basis of an n-dimensional linear space. In a linear space all bases are equivalent and we had no reason there to prefer one basis to another.

There are special, particularly convenient, bases in a Euclidean space called **orthonormal** bases. They play the same role as the Cartesian rectangular basis in analytic geometry. Let us define an orthonormal basis.

Definition. *We say that n elements* $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ of the *n-dimensional Euclidean space E form an* ***orthonormal basis*** *of that space if these elements are pairwise orthogonal and the norm of each of these elements is equal to unity*, i.e. if

$$(\boldsymbol{e}_i, \boldsymbol{e}_k)=\begin{cases}1 \text{ if } i=k,\\ 0 \text{ if } i\neq k.\end{cases} \tag{4.10}$$

To see whether this definition is correct, we must prove that the elements $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ entering into this definition form one of the bases of the n-dimensional space E being considered, and for that purpose, by virtue of Theorem 2.5, it is sufficient to prove that the elements $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ are linearly independent, i.e. that the equality

$$\alpha_1\boldsymbol{e}_1+\alpha_2\boldsymbol{e}_2+\ldots+\alpha_n\boldsymbol{e}_n=\boldsymbol{0} \tag{4.11}$$

is possible only if $\alpha_1=\alpha_2=\ldots=\alpha_n=0$.

Let us prove that this equality holds true. Suppose k is **any** one of the numbers $1, 2, \ldots, n$. Multiplying scalarly equality (4.11) by the element $\boldsymbol{e}_k$ and using the axioms of a scalar product and relations (4.10), we find that $\alpha_k=0$.

Let us now prove the following *fundamental* theorem.

Theorem 4.3. *There is an orthonormal basis in every n-dimensional Euclidean space E.*

Proof. In accordance with the definition of dimensionality there are n linearly independent elements $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n$ in the space E.

We shall prove that it is possible to construct n elements $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ which can be linearly expressed in terms of $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n$ and which form an orthonormal basis (i.e. satisfy relations (4.10)).

Let us prove the possibility of constructing such elements $e_1, e_2, \ldots, e_n$ by induction.

If there is only one element f_1, then, to construct the element e_1 with a norm equal to unity, it is sufficient to normalize the element f_1, i.e. to multiply that element by the number $[\sqrt{(f_1, f_1)}]^{-1}$ inverse to its norm*. As a result we obtain an element $e_1 = [\sqrt{(f_1, f_1)}]^{-1} f_1$ with a norm equal to unity.

Supposing that m is an integer smaller than n, we assume that we have constructed m elements $e_1, e_2, \ldots, e_m$ which can be linearly expressed in terms of $f_1, f_2, \ldots, f_m$, are pairwise orthogonal and have norms equal to unity. Let us prove that we can add to these elements $e_1, e_2, \ldots, e_m$ one more element e_{m+1} which can be linearly expressed in terms of $f_1, f_2, \ldots, f_{m+1}$, is orthogonal to each of the elements $e_1, e_2, \ldots, e_m$ and has a norm equal to unity.

Let us verify that the element e_{m+1} has the form

$$e_{m+1} = \alpha_{m+1}[f_{m+1} - (f_{m+1}, e_m)e_m - (f_{m+1}, e_{m-1})e_{m-1} - \ldots - (f_{m+1}, e_1)e_1], \tag{4.12}$$

where α_{m+1} is some real number.

Indeed, the element e_{m+1} can be linearly expressed in terms of $f_1, f_2, \ldots, f_{m+1}$ (because it can be linearly expressed in terms of $e_1, e_2, \ldots, e_m, f_{m+1}$, and each of the elements $e_1, e_2, \ldots, e_m$ can be linearly expressed in terms of $f_1, f_2, \ldots, f_m$).

It follows immediately that for $\alpha_{m+1} \neq 0$ the element e_{m+1} is obviously **nonzero** (since otherwise a certain linear combination of the linearly independent elements $f_1, f_2, \ldots, f_{m+1}$ would be a zero element in which by virtue of (4.12) the coefficient in f_{m+1} is different from zero.

Next, from the fact that the elements $e_1, e_2, \ldots, e_m$ are pairwise orthogonal and have norms equal to unity and from relation (4.12) it immediately follows that the scalar product (e_{m+1}, e_k) is equal to zero for any number k equal to $1, 2, \ldots, m$.

To complete induction, it remains to prove that the number α_{m+1} may be chosen such that the norm of element (4.12) is equal to unity. We have established above that for $\alpha_{m+1} \neq 0$ the element e_{m+1} and, hence, the element appearing in brackets in (4.12) is nonzero.

Therefore, to normalize the element in brackets, we must take the number α_{m+1} inverse to the positive norm of that element in brackets. In that case, the norm of e_{m+1} is unity. We have thus proved the theorem.

---

* Recall that there cannot be a zero element among the linearly independent elements $f_1, f_2, \ldots, f_n$ and so the norm of f_1 exceeds zero.

The theorem proved leads to the algorithm, carried out step-by-step, for constructing, from the given system of n linearly independent elements $f_1, f_2, \ldots, f_n$ a system of n pairwise orthogonal elements $e_1, e_2, \ldots, e_n$ the norm of each of which is equal to unity:

$$e_1 = \frac{f_1}{\sqrt{(f_1, f_1)}};$$

$$e_2 = \frac{g_2}{\sqrt{(g_2, g_2)}}, \quad \text{where } g_2 = f_2 - (f_2, e_1)\, e_1;$$

$$e_3 = \frac{g_3}{\sqrt{(g_3, g_3)}}, \quad \text{where } g_3 = f_3 - (f_3, e_2)\, e_2 - (f_3, e_1)\, e_1;$$

$$\cdots\cdots\cdots\cdots$$

$$e_n = \frac{g_n}{\sqrt{(g_n, g_n)}}, \quad \text{where } g_n = f_n - (f_n, e_{n-1})\, e_{n-1} - \ldots - (f_n, e_1)\, e_1.$$

This algorithm is known as the **procedure of orthogonalization** of the linearly independent elements $f_1, f_2, \ldots, f_n$.

Remark. There are, of course, many orthonormal bases in every n-dimensional Euclidean space E. Indeed, if, for instance, we use the orthogonalization procedure for constructing an orthonormal basis of the same linear independent elements $f_1, f_2, \ldots, f_n$, then, beginning the orthogonalization process with different elements f_k, we come to different orthonormal bases. In 7.7.2 we shall study the connection between various orthonormal bases of the given Euclidean space E.

We can consider as an example of an orthonormal basis the rectangular Cartesian basis of the Euclidean space of all free vectors or a collection of n elements

$$\begin{aligned} e_1 &= (1, 0, 0, \ldots, 0), \\ e_2 &= (0, 1, 0, \ldots, 0), \\ &\cdots\cdots \\ e_n &= (0, 0, 0, \ldots, 1) \end{aligned}$$

of the Euclidean space E^n of all ordered collections of n real numbers with scalar product (4.2).

4.2.2. The properties of an orthonormal basis. Suppose $e_1, e_2, \ldots, e_n$ is an arbitrary orthonormal basis of the n-dimensional Euclidean space E, and x and y are two arbitrary elements of that space. Let us find the expression for the scalar product (x, y) of these elements in terms of their coordinates relative to the basis $e_1, e_2, \ldots, e_n$.

We designate the coordinates of the elements x and y relative to the basis $e_1, e_2, \ldots, e_n$ as $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ respective-

ly, i.e. assume that $\boldsymbol{x}=x_1\boldsymbol{e}_1+x_2\boldsymbol{e}_2+\ldots+x_n\boldsymbol{e}_n$, $\boldsymbol{y}=y_1\boldsymbol{e}_1+y_2\boldsymbol{e}_2+\ldots$ $\ldots+y_n\boldsymbol{e}_n$. Then

$$(\boldsymbol{x},\ \boldsymbol{y})=(x_1\boldsymbol{e}_1+x_2\boldsymbol{e}_2+\ldots+x_n\boldsymbol{e}_n,\ y_1\boldsymbol{e}_1+y_2\boldsymbol{e}_2+\ldots+y_n\boldsymbol{e}_n).$$

By virtue of the axioms of a scalar product and relations (4.10) it follows from the last equation that

$$(\boldsymbol{x},\ \boldsymbol{y})=\left(\sum_{i=1}^{n} x_i\boldsymbol{e}_i,\ \sum_{k=1}^{n} y_k\boldsymbol{e}_k\right)=\sum_{i=1}^{n}\sum_{k=1}^{n} x_iy_k(\boldsymbol{e}_i,\ \boldsymbol{e}_k)$$
$$=x_1y_1+x_2y_2+\ldots+x_ny_n.$$

Thus, in the final analysis we have

$$(\boldsymbol{x},\ \boldsymbol{y})=x_1y_1+x_2y_2+\ldots+x_ny_n. \tag{4.13}$$

Thus, *in an orthonormal basis the scalar product of any two elements is equal to the sum of the products of the respective coordinates of those elements.*

Let us consider now an arbitrary (nonorthonormal) basis $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots$ $\ldots, \boldsymbol{f}_n$ in the n-dimensional Euclidean space E and find the expression of the scalar product of two arbitrary elements $\boldsymbol{x}$ and $\boldsymbol{y}$ in terms of their coordinates with respect to the indicated basis.

We designate the coordinates of the elements $\boldsymbol{x}$ and $\boldsymbol{y}$ with respect to the basis $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n$ as $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ respectively, i.e. assume that

$$\boldsymbol{x}=x_1\boldsymbol{f}_1+x_2\boldsymbol{f}_2+\ldots+x_n\boldsymbol{f}_n,\ \boldsymbol{y}=y_1\boldsymbol{f}_1+y_2\boldsymbol{f}_2+\ldots+y_n\boldsymbol{f}_n.$$

Applying the axioms of a scalar product, we get

$$(\boldsymbol{x},\ \boldsymbol{y})=(x_1\boldsymbol{f}_1+x_2\boldsymbol{f}_2+\ldots+x_n\boldsymbol{f}_n,$$
$$y_1\boldsymbol{f}_1+y_2\boldsymbol{f}_2+\ldots+y_n\boldsymbol{f}_n)=\left(\sum_{i=1}^{n} x_i\boldsymbol{f}_i,\ \sum_{k=1}^{n} y_k\boldsymbol{f}_k\right)$$
$$=\sum_{i=1}^{n}\sum_{k=1}^{n} x_iy_k(\boldsymbol{f}_i,\ \boldsymbol{f}_k).$$

Thus, *in an arbitrary basis* $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n$, *the scalar product of any two elements* $\boldsymbol{x}=x_1\boldsymbol{f}_1+x_2\boldsymbol{f}_2+\ldots+x_n\boldsymbol{f}_n$ and $\boldsymbol{y}=y_1\boldsymbol{f}_1+y_2\boldsymbol{f}_2+\ldots$ $\ldots+y_n\boldsymbol{f}_n$ *is defined by the equation*

$$(\boldsymbol{x},\ \boldsymbol{y})=\sum_{i=1}^{n}\sum_{k=1}^{n} a_{ik}x_iy_k,$$

in which the matrix $\|a_{ik}\|$ $(i=1, 2, \ldots, n;\ k=1, 2, \ldots, n)$ *has the elements* $a_{ik}=(\boldsymbol{f}_i,\ \boldsymbol{f}_k)$.

The last assertion leads to the following result: *for the scalar product of any two elements in the given basis* $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n$ *of the*

Euclidean space E to be equal to the sum of products of the respective coordinates of those elements, it is necessary and sufficient that the basis $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n$ *be orthonormal.*

In fact, expression (4.14) passes into (4.13) if and only if the matrix $\| a_{ik} \|$ with the elements $a_{ik}=(\boldsymbol{f}_i, \boldsymbol{f}_k)$ is an identity matrix, i.e. if and only if the relations

$$(\boldsymbol{f}_i, \boldsymbol{f}_k)=\begin{cases} 1 \text{ for } i=k, \\ 0 \text{ for } i\neq k, \end{cases}$$

establishing the orthonormality of the basis $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n$, are satisfied.

Let us continue with the discussion of the arbitrary orthonormal basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ of the n-dimensional Euclidean space E. We shall elucidate the sense of the coordinates of the arbitrary element $\boldsymbol{x}$ with respect to the indicated basis.

We designate the coordinates of the element $\boldsymbol{x}$ with respect to the basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ as $x_1, x_2, \ldots, x_n$, i.e. assume

$$\boldsymbol{x}=x_1\boldsymbol{e}_1+x_2\boldsymbol{e}_2+\ldots+x_n\boldsymbol{e}_n. \tag{4.15}$$

Next we designate as k any of the numbers $1, 2, \ldots, n$ and multiply both sides of (4.5) scalarly by the element $\boldsymbol{e}_k$. On the basis of the axioms of a scalar product and relations (4.10), we obtain

$$(\boldsymbol{x}, \boldsymbol{e}_k)=\left(\sum_{i=1}^{n} x_i\boldsymbol{e}_i, \boldsymbol{e}_k\right)=\sum_{i=1}^{n} x_i(\boldsymbol{e}_i, \boldsymbol{e}_k)=x_k.$$

Thus, *the coordinates of an arbitrary element relative to an orthonormal basis are equal to the scalar products of that element by the respective components.*

Since the scalar product of the arbitrary element $\boldsymbol{x}$ by the element $\boldsymbol{e}$ with a norm equal to unity is naturally called **the projection of the element $\boldsymbol{x}$ by the element $\boldsymbol{e}$,** we may say that *the coordinates of an arbitrary element relative to an orthonormal basis are equal to the projections of that element by the respective components.*

Thus an arbitrary orthonormal basis possesses properties which are quite analogous to the properties of a rectangular Cartesian basis.

4.2.3. Decomposition of an n-dimensional Euclidean space in a direct sum of its subspace and its orthogonal complement. Suppose G is an arbitrary subspace of an n-dimensional Euclidean space E.

The collection F of all elements $\boldsymbol{y}$ of the space E which are orthogonal to every element $\boldsymbol{x}$ of the subspace G is known as an ***orthogonal complement*** *of the subspace G.*

Note that the orthogonal complement F is itself a subspace of E (since it evidently follows from the orthogonality of every element

y_1 and y_2 to the element x that any linear combination of the elements y_1 and y_2 is orthogonal to the element x).

Let us prove that *every n-dimensional Euclidean space E is a direct sum of its arbitrary subspace G and its orthogonal complement F.*

We choose in G an arbitrary orthonormal basis $e_1, e_2, \ldots, e_k$. By virtue of what was proved in 2.3.1, this basis can be complemented by the elements $f_{k+1}, \ldots, f_n$ of the space E to complete the basis in the whole E. Carrying out the **process of orthogonalization** of elements $e_1, \ldots, e_k, f_{k+1}, \ldots, f_n$, we get an orthonormal basis $e_1, \ldots, e_k, e_{k+1}, \ldots, e_n$ of the whole space E. Expressing the arbitrary element x of the space E in terms of these components, i.e. representing it as $x = x_1e_1 + \ldots + x_ke_k + x_{k+1}e_{k+1} + \ldots + x_ne_n$, we find that the element x can be uniquely represented in the form $x = x' + x''$, where $x' = x_1e_1 + \ldots + x_ke_k$ is a definite element of G and $x'' = x_{k+1}e_{k+1} + \ldots + x_ne_n$ is a definite element of the orthogonal complement F (each element $e_{k+1}, \ldots, e_n$ is orthogonal to any of the elements $e_1, \ldots, e_k$ and is, therefore, orthogonal to any element of G, the linear combination $x_{k+1}e_{k+1} + \ldots + x_ne_n$ is, therefore, also orthogonal to any element of G, i.e. is a definite element of F).

4.2.4. The isomorphism of *n*-dimensional Euclidean spaces. We shall show here that various Euclidean spaces of the same dimensionality n do not essentially differ from one another in the sense of the properties connected with the operations introduced in those spaces.

Since only operations of addition of elements, multiplication of elements by scalars and a scalar multiplication of elements were introduced in Euclidean spaces, it is natural to formulate the following definition.

Definition. *Two Euclidean spaces E and E′ are said to be* ***isomorphic*** *if a one-to-one correspondence can be established between the elements of these spaces such that if the elements* x *and* y *of the space E are associated with the elements* x' *and* y' *of the space E′, then the element* $x+y$ *is associated with the element* $x'+y'$, *the element* λx *(for any real* λ*) is associated with the element* $\lambda x'$ *and the scalar product* (x, y) *is equal to the scalar product* (x', y').

Thus, the Euclidean spaces E and E' are isomorphic if they are isomorphic as linear spaces* and if that isomorphism retains the value of the scalar product of the corresponding pairs of elements.

Theorem 4.4. *All Euclidean spaces of the same dimension n are isomorphic to one another.*

Proof. It is sufficient to prove that any n-dimensional Euclidean space E' is isomorphic to the Euclidean space E^n of ordered collections of n real numbers with scalar product (4.2). According to Theo-

* See 2.2.4.

rem 4.3, there is an orthonormal basis $e'_1, e'_2, \ldots, e'_n$ in the Euclidean space E'. We put every element $x' = x_1e'_1 + x_2e'_2 + \ldots + x_ne'_n$ of the space E' into correspondence with n real numbers $x_1, x_2, \ldots, x_n$, i.e. a definite element $x = (x_1, x_2, \ldots, x_n)$ of the space E^n.

The correspondence we have established is one-to-one. In addition, it follows from Theorem 2.4 that if the elements $x' = (x_1, x_2, \ldots, x_n)$ and $y' = (y_1, y_2, \ldots, y_n)$ of the space E'* are associated with the respective elements $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ of the space E^n, then the element $x' + y'$ is associated with the element $x + y$ and the element $\lambda x'$ is associated with the element λx.

It remains to prove that the value of the scalar product is retained for the respective pairs of elements x', y' and x, y.

Because of the orthonormality of the basis $e'_1, e'_2, \ldots, e'_n$ and according to formula (4.13), we have $(x', y') = x_1y_1 + x_2y_2, \ldots, x_ny_n$. On the other hand, by virtue of formula (4.2) defining the scalar product in the space E^n, $(x, y) = x_1y_1 + x_2y_2 + \ldots + x_ny_n$. And this completes the proof of the theorem.

This theorem enables us to assert that if in any specific n-dimensional Euclidean space E' the theorem formulated in terms of operations of addition, multiplication by scalars and scalar multiplication of elements has been proved, then that theorem is valid in an arbitrary n-dimensional Euclidean space E as well.

4.3. A Complex Euclidean Space

4.3.1. Definition of a complex Euclidean space. We indicated at the end of 2.1.1 that if in the definition of a linear space we take the numbers λ, μ, ... not from the set of real numbers but from the set of all complex numbers, then we arrive at the concept of a complex linear space.

A complex Euclidean space which is of a fundamental importance in the theory of nonself-adjoint linear transformations is constructed on the basis of complex linear spaces.

To introduce a complex linear space we must first introduce the notion of a scalar product of two elements of a complex linear space subordinated to the corresponding four axioms.

Definition. *The complex linear space R is called a* ***complex Euclidean space*** *if the following two requirements are met:*

I. *There is a rule which puts any two elements x and y of that space into correspondence with a* ***complex*** *number called the* ***scalar product*** *of those elements and symbolized as (x, y).*

---

* The coordinates of these elements are taken with respect to the basis $e'_1, e'_2, \ldots, e'_n$.

II. *The indicated rule is subordinated to the following four axioms*

1°. $(x, y)=\overline{(y, x)}$*.

2°. $(x_1+x_2, y)=(x_1, y)+(x_2, y)$.

3°. $(\lambda x, y)=\lambda(x, y)$.

4°. (x, x) *is a real nonnegative number vanishing only when* x *is a zero element***.

The following two relations are logical consequences of axioms 1°-3°:

$$(x, \lambda y)=\bar{\lambda}(x, y), \ (x, y_1+y_2)=(x, x_1)+(x, y_2).$$

Indeed, we infer from axioms 1° and 3° that

$$(x, \lambda y)=\overline{(\lambda y, x)}=\bar{\lambda}\overline{(y, x)}=\bar{\lambda}(x, y),$$

and from axioms 1° and 2° we find that

$$(x, y_1+y_2)=\overline{(y_1+y_2, x)}=\overline{(y_1 x)}+\overline{(y_2, x)}=(x, y_1)+(x, y_2).$$

Here are some examples of specific complex Euclidean spaces.

Example 1. Let us consider the collection $C^*[a, b]$ of all functions $z=z(t)$ defined for the values of t from the interval $a\leqslant t\leqslant b$ and assuming complex values $z(t)=x(t)+iy(t)$ such that the real functions $x(t)$ and $y(t)$ are continuous on that interval. The operations of addition of those functions and their multiplication by complex numbers will be taken from the analysis. The scalar product of any two functions of this kind is defined by the relation $(z_1(t), z_2(t))=\int_a^b z_1(t)\overline{z_2(t)}\,dt$.

It is easy to verify all the axioms 1°-4° for a scalar product thus defined, from which it follows that the collection being considered is a complex Euclidean space.

Example 2. Let us consider the complex linear space A_*^n whose elements are ordered collections of n **complex** numbers $x_1, x_2, \ldots, x_n$ with the definitions of operations of addition of elements and their multiplication by scalars as in the case of the real linear space A^n.

The scalar product of any two elements $x=(x_1, x_2, \ldots, x_n)$ and

---

* Here and in what follows we use the symbol $\bar{\alpha}$ to denote a number which is a complex conjugate of α.

** Axiom 1° differs from the corresponding axiom 1° of a real Euclidean space. It is easy to verify that in passing to a complex space it is impossible to retain unchanged all the three axioms 1°, 3° and 4° of a real scalar product. Indeed, with the axioms $(x, y)=(y, x)$ and $(\lambda x, y)=\lambda(x, y)$, we would get $(x, \lambda y)=(\lambda y, x)=\lambda(y, x)=\lambda(x, y)$. But then it would turn out that $(\lambda x, \lambda x)=\lambda^2(x, y)$ and, hence, for $\lambda=i$ we would find that $(ix, ix)=-(x, x)$ and that would contradict axiom 4° and the nonnegativity of (y, y) for any element y.

$y=(y_1, y_2, \ldots, y_n)$ is defined by the relation

$$(x, y)=x_1\bar{y}_1+x_2\bar{y}_2+\ldots+x_n\bar{y}_n. \tag{4.16}$$

It is easy to verify the validity of axioms 1°-3° for the scalar product thus defined. The validity of axiom 4° follows from the relation

$$(x, x)=x_1\bar{x}_1+x_2\bar{x}_2+\ldots+x_n\bar{x}_n=|x_1|^2+|x_2|^2+\ldots+|x_n|^2.$$

This means that the space A_*^n with the scalar product (4.16) is a complex Euclidean space.

Example 3. We can introduce the scalar product in the same complex linear space A_*^n not by relation (4.16) but by a more general relation*

$$(x, y)=\sum_{i=1}^{n}\sum_{k=1}^{n} a_{ik}x_i\bar{y}_k, \tag{4.17}$$

in which $\| a_{ik} \|$ is an arbitrary matrix consisting of complex numbers a_{ik} which satisfy the condition $a_{ik}=\bar{a}_{ki}$, such that the quadratic form $\sum_{i=1}^{n}\sum_{k=1}^{n} a_{ik}x_i\bar{x}_k$ assumes nonnegative real values for all complex numbers $x_1, x_2, \ldots, x_n$ and vanishes only under the condition $|x_1|^2+|x_2|^2+\ldots+|x_n|^2=0$.

We invite the reader to verify that the scalar product thus defined satisfies axioms 1°-4°.

4.3.2. The Cauchy-Bunyakovsky inequality. The concept of a norm. *Let us prove that the* ***Cauchy-Bunyakovsky inequality*** *holds true for any two elements* x *and* y *of an arbitrary complex Euclidean space***

$$|(x, y)|^2 \leqslant (x, x)(y, y). \tag{4.18}$$

On the basis of axiom 4° the inequality

$$(\lambda x-y, \lambda x-y)\geqslant 0 \tag{4.19}$$

is valid for any complex number λ. Since by virtue of axioms 1°-3° and their logical consequences we have

$$\begin{aligned}(\lambda x-y, \lambda x-y)&=\lambda\bar{\lambda}(x, x)-\lambda(x, y)-\bar{\lambda}(y, x)+(y, y)\\ &=|\lambda|^2(x, x)-\lambda(x, y)-\bar{\lambda}\overline{(x, y)}+(y, y),\end{aligned}$$

* (1.17) passes into (4.16) when the matrix $\| a_{ik} \|$ is an identity matrix.

** Since (x, y) is, in general, a complex number, we cannot write the Cauchy-Bunyakovsky inequality in form (4.6).

inequality (4.19) assumes the form

$$|\lambda|^2(\boldsymbol{x}, \boldsymbol{x})-\lambda(\boldsymbol{x}, \boldsymbol{y})-\bar{\lambda}\overline{(\boldsymbol{x}, \boldsymbol{y})}+(\boldsymbol{y}, \boldsymbol{y})\geqslant 0. \tag{4.20}$$

Let us denote the **argument** of the complex number $(\boldsymbol{x}, \boldsymbol{y})$ by φ and represent this number in **trigonometric** form*

$$(\boldsymbol{x}, \boldsymbol{y})=|(\boldsymbol{x}, \boldsymbol{y})|(\cos\varphi+i\sin\varphi). \tag{4.21}$$

Now we set the complex number

$$\lambda=t(\cos\varphi-i\sin\varphi), \tag{4.22}$$

where t is an **arbitrary real number.** It is evident from relations (4.21) and (4.22) that $|\lambda|=|t|$, $\lambda(\boldsymbol{x}, \boldsymbol{y})=\bar{\lambda}\overline{(\boldsymbol{x}, \boldsymbol{y})}=t|(\boldsymbol{x}, \boldsymbol{y})|$. Therefore, for λ we have chosen, inequality (4.20) turns into an inequality

$$t^2(\boldsymbol{x}, \boldsymbol{x})-2t|(\boldsymbol{x}, \boldsymbol{y})|+(\boldsymbol{y}, \boldsymbol{y})\geqslant 0, \tag{4.23}$$

which is valid for any real t. The necessary and sufficient condition for the quadratic trinomial on the left-hand side of (4.23) to be nonnegative is the nonpositivity of its discriminant, i.e. the inequality $|(\boldsymbol{x}, \boldsymbol{y})|^2-(\boldsymbol{x}, \boldsymbol{x})(\boldsymbol{y}, \boldsymbol{y})\leqslant 0$ which is equivalent to inequality (4.18).

We can establish with the aid of the Cauchy-Bunyakovsky inequality (4.18) and arguments completely analogous to the proof of Theorem 4.2 that *every complex Euclidean space is normed if the norm of any element* $\boldsymbol{x}$ *in it is defined by the relation*

$$\|\boldsymbol{x}\|=\sqrt{(\boldsymbol{x}, \boldsymbol{x})}. \tag{4.24}$$

In particular, *in any complex Euclidean space with the norm defined by relation* (4.24) *there holds a* ***triangle inequality*** $\|\boldsymbol{x}+\boldsymbol{y}\|\leqslant\|\boldsymbol{x}\|+\|\boldsymbol{y}\|$.

Remark. It should be emphasized that the notion of the angle φ between two arbitrary elements $\boldsymbol{x}$ and $\boldsymbol{y}$ introduced for a **real** Euclidean space loses sense for a **complex** Euclidean space (because the scalar product $(\boldsymbol{x}, \boldsymbol{y})$ is, in general, a complex number).

4.3.3. An orthonormal basis and its properties. The elements $\boldsymbol{x}$ and $\boldsymbol{y}$ of an arbitrary complex Euclidean space are said to be **orthogonal** if the scalar product $(\boldsymbol{x}, \boldsymbol{y})$ of these elements is equal to zero. The orthonormal basis of an n-dimensional complex Euclidean space is the collection of its elements $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ satisfying the relations

$$(\boldsymbol{e}_i, \boldsymbol{e}_k)=\begin{cases} 1 \text{ for } i=k, \\ 0 \text{ for } i\neq k \end{cases} \tag{4.25}$$

(i.e. pairwise orthogonal and having norms equal to unity).

* You can find concepts of an argument and a trigonometric form of a complex number in 7.1 of *Fundamentals of Mathematical Analysis,* Part I, by Ilyin and Poznyak.

We can prove in the same way as in 2.1 that these elements are linearly independent and, therefore, form a basis.

By a complete analogy with the proof of Theorem 4.3 (i.e. with the aid of the orthogonalization procedure) we establish the existence of an orthonormal basis in an arbitrary n-dimensional complex Euclidean space.

Let us express the scalar product of two arbitrary elements $\boldsymbol{x}$ and $\boldsymbol{y}$ of an n-dimensional complex Euclidean space in terms of their coordinates $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$ relative to the orthonormal basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$. Since

$$\boldsymbol{x}=x_1\boldsymbol{e}_1+x_2\boldsymbol{e}_2+\ldots+x_n\boldsymbol{e}_n,\quad \boldsymbol{y}=y_1\boldsymbol{e}_1+y_2\boldsymbol{e}_2+\ldots+y_n\boldsymbol{e}_n,$$

we obtain, by virtue of axioms 1°-4° and relations (4.25),

$$(\boldsymbol{x}, \boldsymbol{y})=\left(\sum_{i=1}^{n} x_i\boldsymbol{e}_i, \sum_{k=1}^{n} y_k\boldsymbol{e}_k\right)$$

$$=\sum_{i=1}^{n}\sum_{k=1}^{n} x_i\bar{y}_k(\boldsymbol{e}_i, \boldsymbol{e}_k)=x_1\bar{y}_1+x_2\bar{y}_2+\ldots+x_n\bar{y}_n.$$

Next we express the coordinates $x_1, x_2, \ldots, x_n$ of the arbitrary element $\boldsymbol{x}$ relative to the orthonormal basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$.

Multiplying the expression of this element in terms of the components $\boldsymbol{x}=x_1\boldsymbol{e}_1+x_2\boldsymbol{e}_2+\ldots+x_n\boldsymbol{e}_n$ scalarly by $\boldsymbol{e}_k$ and using relations (4.25), we obtain (for any k equal to 1, 2, ..., n)

$$(\boldsymbol{x}, \boldsymbol{e}_k)=\left(\sum_{i=1}^{n} x_i\boldsymbol{e}_i, \boldsymbol{e}_k\right)=\sum_{i=1}^{n} x_i(\boldsymbol{e}_i, \boldsymbol{e}_k)=x_k.$$

Thus, as in the case of a real Euclidean space, *the coordinates of the arbitrary element* $\boldsymbol{x}$ *with respect to an orthonormal basis are equal to the scalar products of that element by the corresponding base elements.*

By a complete analogy with the proof of Theorem 4.4 we establish that *all complex Euclidean spaces of the same dimension* n *are isomorphic to each other.*

4.4. The Regularization Method for Seeking a Normal Solution of a Linear system

We return to the general linear system of m equations in n unknowns of the form (3.1). We write this system in short, in matrix form*

$$AX=B. \tag{4.26}$$

* See formula 3.6 in the preceding chapter.

Recall that in this notation the symbol A denotes the matrix $A = \| a_{ij} \|$ $(i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n)$, and the symbols X and B denote the columns (or vectors) of the form

$$X = \begin{Vmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{Vmatrix}, \quad B = \begin{Vmatrix} b_1 \\ b_2 \\ \cdots \\ b_m \end{Vmatrix}$$

the first of which must be found and the second is given.

We shall discuss the case when the values of the elements of the matrix A and of the column of constant terms B are given only approximately*. Then it is natural to speak only of approximate values of the sought-for column X. The algorithms for calculating the column of solutions X presented in the previous chapter and based on Cramer's rule may lead to large errors in this case and lose practical significance**.

We shall give here an algorithm belonging to Tikhonov and making it possible to find the so-called **normal** (i.e. the closest to the origin) solution of X with an accuracy corresponding to the accuracy of specifying the elements of the matrix A and the column B***.

Let us introduce the so-called *spherical norms* of the columns B and X of the matrix A setting them equal to

$$\| B \| = \sqrt{\sum_{i=1}^{m} b_i^2}, \quad \| X \| = \sqrt{\sum_{j=1}^{n} x_j^2},$$
$$\| A \| = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}. \tag{4.27}$$

Note that the norms of the columns B and X are defined as ordinary norms of vectors, i.e. the elements of the spaces E^m and E^n respectively. The norm of the matrix A is defined coherently with the norm

* Such a situation occurs when these values are obtained as a result of physical measurements or if in the process of calculations we have to round off the indicated values of a certain digit.

** This especially refers to the case of the so-called ill-conditioned matrices (for which "small" changes in the elements of the matrix of the base minor lead to "great" changes in the elements of the inverse matrix.

*** See A. N. Tikhonov, "On Incorrect Problems of Linear Algebra and the Stable Method of Their Solution", Doklady Akademii Nauk SSSR, Vol. 163, 3 (1965), pp. 591-594 (in Russian).

of the n-dimensional column X in the sense that the norm of the m-dimensional column AX, equal to the product of the matrix A by the column X, satisfies the condition*

$$\| AX \| \leqslant \| A \| \ \| X \|. \tag{4.28}$$

We assume that instead of exact values of the elements of the matrix $A = \| a_{ij} \|$ and the column of the right-hand sides $B = \| b_i \|$ we are given approximate values $\tilde{A} = \| \tilde{a}_{ij} \|$, $\tilde{B} = \| \tilde{b}_i \|$.

We call the matrix $\tilde{A}$ (the column $\tilde{B}$) the **δth approximation** of the matrix A (column B) if the following inequality holds true:

$$\| A - \tilde{A} \| = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} (a_{ij} - \tilde{a}_{ij})^2} < \delta$$

$$\left(\| B - \tilde{B} \| = \sqrt{\sum_{i=1}^{m} (b_i - \tilde{b}_i)^2} \leqslant \delta \right). \tag{4.29}$$

The **normal solution** of the consistent system (4.26) is assumed to be its solution $X^0 = \begin{Vmatrix} x_1^0 \\ x_2^0 \\ \cdots \\ x_n^0 \end{Vmatrix}$ whose norm $\| X^0 \|$ is the least of the norms $\| X \|$ of all the solutions X of that system. Note that every consistent system (4.26) (an indeterminate system inclusive) possesses a unique normal solution.

We introduce into consideration the following function of n variables $x_1, x_2, \ldots, x_n$ or of one column $X = \begin{Vmatrix} x_1 \\ \cdots \\ x_n \end{Vmatrix}$

$$F^\alpha(x_1, \ldots, x_n \ \tilde{A}, \tilde{B}) = F^\alpha(X, \tilde{A}, \tilde{B}) = \| \tilde{A}X - \tilde{B} \|^2 + \alpha \| X \|^2 \tag{4.30}$$

dependent on the elements of the matrix $\tilde{A}$ and the column $\tilde{B}$ as

---

* Indeed, using the definition of the product of a matrix by a column, relations (4.27) and the Cauchy-Bunyakovsky inequality for the elements of the Euclidean space E^n, we get

$$\| AX \|^2 = \sum_{i=1}^{m} \left[\sum_{j=1}^{n} a_{ij} x_j \right]^2 \leqslant \sum_{i=1}^{m} \left[\sum_{j=1}^{n} a_{ij}^2 \cdot \sum_{j=1}^{n} x_j^2 \right]$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2 \cdot \sum_{j=1}^{n} x_j^2 = \| A \|^2 \| X \|^2.$$

parameters and also on a certain numerical parameter α. This function can be written out as

$$F^{\alpha}=(X,\ \tilde{A},\ \tilde{B})=\sum_{i=1}^{m}\left[\sum_{j=1}^{n}\tilde{a}_{ij}x_j-\tilde{b}_i\right]^2+\alpha\sum_{j=1}^{n}x_j^2. \qquad (4.30')$$

Actually $F^{\alpha}(X,\ \tilde{A},\ \tilde{B})$ is a function of the elements of X of the Euclidean space E^n of the n-dimensional columns. A function of this kind, whose argument are the elements of a certain linear space, is customarily called a **functional***.

It is easy to verify that *for any fixed* $\alpha>0$, *the nonnegative functional* (4.30′) *attains its minimum* (in the whole space E^n) *value at the only point* $X^{\alpha}=\begin{Vmatrix} x_1^{\alpha} \\ \cdots \\ x_n^{\alpha} \end{Vmatrix}$ *of the space* E^n.

Indeed, differentiating function (4.30′) twice, we get

$$\frac{\partial^2 F^{\alpha}}{\partial x_k \partial x_l}=2\sum_{i=1}^{m}\tilde{a}_{ik}\tilde{a}_{il}+2\alpha\delta_{kl}, \quad \text{where } \delta_{kl}=\begin{cases} 1 \text{ for } k=l, \\ 0 \text{ for } k\neq l. \end{cases}$$

Consequently, the second differential of the function F^{α} has the form

$$d^2F^{\alpha}=\sum_{k=1}^{n}\sum_{l=1}^{n}\left[\sum_{i=1}^{m}\tilde{a}_{ik}\tilde{a}_{il}\right]dx_k dx_l+\alpha\sum_{k=1}^{n}\sum_{l=1}^{n}\delta_{kl}dx_k dx_l$$

$$=\sum_{i=1}^{m}\left[\sum_{k=1}^{n}\tilde{a}_{ik}dx_k\right]^2+\alpha\sum_{k=1}^{n}(dx_k)^2.$$

This equation yields an estimation $d^2F^{\alpha}\geqslant\alpha\sum_{k=1}^{n}(dx_k)^2$ signifying that the function F^{α} is **strictly convex downwards.** In addition, $F^{\alpha}\to+\infty$ for $\|X\|=\sqrt{\sum_{k=1}^{n}x_k^2}\to\infty$. From this it obviously follows that F^{α} has a point of minimum X^{α} and that this point is unique**.

The methods of seeking the minimum value of the functionals of the form (4.30) are well elaborated***.

---

* Functionals are systematically studied in the next chapter.

** See, in particular, *The Fundamentals of Mathematical Analysis* by Ilyin and Poznyak, Part 1, 14.7.

*** Ditto.

Let us prove the following fundamental theorem reducing the problem of an approximate seeking the normal solution of system (4.26) to finding the element $X^\alpha = \begin{Vmatrix} x_1^\alpha \\ \cdots \\ x_n^\alpha \end{Vmatrix}$ on which functional (4.30) attains its minimum value.

Theorem of Tikhonov. *Suppose the matrix A and the column B satisfy the conditions ensuring the consistency of system* (4.26), $X^0 = \begin{Vmatrix} x_1^0 \\ \cdots \\ x_n^0 \end{Vmatrix}$ *is a normal solution of that system, $\tilde{A}$ is the δ-approximation of the matrix A, $\tilde{B}$ is the δ-approximation of the column B, $\varepsilon(\delta)$ and $\alpha(\delta)$ are some increasing functions δ tending to zero as $\delta \to 0+0$ and such that $\delta^2 \leqslant \varepsilon(\delta)\,\alpha(\delta)$. Then, for any $\varepsilon > 0$ there is a positive number $\delta_0 = \delta_0(\varepsilon, \| X^0 \|)$ such that for any $\delta < \delta_0(\varepsilon, \| X^0 \|)$ and any α satisfying the condition*

$$\frac{1}{\varepsilon(\delta)}\,\delta^2 \leqslant \alpha \leqslant \alpha(\delta) \tag{4.31}$$

the element $X^\alpha = \begin{Vmatrix} x_1^\alpha \\ \cdots \\ x_n^\alpha \end{Vmatrix}$ furnishing functional (4.30) *with a minimum satisfies the inequality*

$$\| X^\alpha - X^0 \| \leqslant \varepsilon. \tag{4.32}$$

Proof. Let us consider in the linear space E^m a subset $\{U_{\tilde{A}}\}$ of all elements $U = \begin{Vmatrix} u_1 \\ \cdots \\ u_m \end{Vmatrix}$ which can be represented as $U = \tilde{A}X$, where $X = \begin{Vmatrix} x_1 \\ \cdots \\ x_n \end{Vmatrix}$ is an arbitrary element of the space E^n. It is quite evident that the subset $\{U_{\tilde{A}}\}$ is a linear space and is, therefore, a subspace of E^m. We denote by $\{V_{\tilde{A}}\}$ an orthogonal complement of $\{U_{\tilde{A}}\}$ (to complete E^m) and decompose E^m in a direct sum of the subspaces $\{U_{\tilde{A}}\}$ and $\{V_{\tilde{A}}\}$*. Assume that $\tilde{B}_{\tilde{A}}$ denotes a projection

* See 4.2.3.

of the column $\tilde{B}$ onto the subspace $\{U_{\tilde{A}}\}$ so that $\tilde{B}=\tilde{B}_{\tilde{A}}+(\tilde{B}-\tilde{B}_{\tilde{A}})$, where $(\tilde{B}-\tilde{B}_{\tilde{A}})$ is an element of $\{V_{\tilde{A}}\}$. Then, since for any element X of the space E^n the column $\tilde{A}X$ is an element of $\{U_{\tilde{A}}\}$, we get the following decomposition:

$$\tilde{A}X-\tilde{B}=(\tilde{A}X-\tilde{B}_{\tilde{A}})+(\tilde{B}_{\tilde{A}}-\tilde{B}),$$

in which the elements $(\tilde{A}X-\tilde{B}_{\tilde{A}})$ and $(\tilde{B}_{\tilde{A}}-\tilde{B})$ are orthogonal to each other and belong to $\{U_{\tilde{A}}\}$ and $\{V_{\tilde{A}}\}$ respectively.

Applying Pythagoras' theorem (see 4.1.2), we get (for any element X of the space E^n)

$$\| \tilde{A}X-\tilde{B} \|^2=\| \tilde{A}X-\tilde{B}_{\tilde{A}} \|^2+\| \tilde{B}-\tilde{B}_{\tilde{A}} \|^2, \tag{4.33}$$

which yields, in particular, an inequality

$$\| \tilde{B}-\tilde{B}_{\tilde{A}} \| \leqslant \| \tilde{A}X-\tilde{B} \| \tag{4.34}$$

which is also valid for any element X of the space E^n.

From (4.33) and (4.30) we find that for any X from E^n

$$F^{\alpha}(X, \tilde{A}, \tilde{B})=\| \tilde{B}-\tilde{B}_{\tilde{A}} \|^2+F^{\alpha}(X, \tilde{A}, \tilde{B}_{\tilde{A}}), \tag{4.35}$$

i.e. the functionals on the left-hand and right-hand sides of (4.35) *have a common element X^{α} furnishing them with a minimum.*

Let us find now for any α satisfying conditions (4.31) the inequality

$$F^{\alpha}(X^{\alpha}, \tilde{A}, \tilde{B}_{\tilde{A}}) \leqslant \alpha\varepsilon(\delta)\, C^2+\alpha \| X^0 \| \tag{4.36}$$

in which C denotes the quantity $C=2(1+\| X^0 \|)$, and X^0 is a normal solution of system (4.26).

Since the column X^{α} furnishes the functional appearing on the right-hand side of (4.35) with a minimum, we have

$$F^{\alpha}(X^{\alpha}, \tilde{A}, \tilde{B}_{\tilde{A}}) \leqslant F^{\alpha}(X^0, \tilde{A}, \tilde{B}_{\tilde{A}})=\| \tilde{A}X_0-\tilde{B}_{\tilde{A}} \|^2+\alpha \| X^0 \|^2. \tag{4.37}$$

Using the relation $AX^0=B$ and the triangle inequality, we obtain $\| \tilde{A}X^0-\tilde{B}_{\tilde{A}} \| \leqslant \| \tilde{A}X^0-AX^0 \|+\| B-\tilde{B} \|+\| \tilde{B}-\tilde{B}_{\tilde{A}} \|$. On the right-hand side of the last inequality we use relations (4.28) and (4.29) and also inequality (4.34) taken for $X=X^0$. We get

$$\| \tilde{A}X^0-\tilde{B}_{\tilde{A}} \| \leqslant \delta \| X^0 \|+\delta+\| \tilde{B}+\tilde{A}X^0 \|. \tag{4.38}$$

Taking into account once again that $AX^0=B$ and using the triangle inequality and relations (4.28) and (4.29), we find that

$$\| \tilde{B}-\tilde{A}X^0 \| \leqslant \| \tilde{B}-B \|+\| AX^0-\tilde{A}X^0 \| \leqslant \delta+\delta\cdot \| X^0 \|. \tag{4.39}$$

It follows from (4.38) and (4.39) that

$$\| \tilde{A}X^0 - \tilde{B}_{\tilde{A}} \| \leqslant 2\delta (1 + \| X^0 \|) = C\delta, \tag{4.40}$$

where $C = 2(1 + \| X^0 \|)$.

To complete the proof of estimate (4.36), it remains to substitute (4.40) into (4.37) and use inequality (4.31).

Since it immediately follows from the definition of the functional F^α that $\alpha \cdot \| X^\alpha \|^2 \leqslant F^\alpha (X^\alpha, \tilde{A}, \tilde{B}_{\tilde{A}})$, inequality (4.36) we have proved also yields an inequality

$$\| X^\alpha \| \leqslant \| X^0 \| + \varepsilon_1 (\delta), \tag{4.41}$$

in which $\varepsilon_1 (\delta) \to 0$ as $\delta \to 0+0$. It follows from (4.41) that for all sufficiently small δ *the set* $\{ X^\alpha \}$ *of the points* X^α *of the space* E^n *is* ***bounded.***

It is now easy to prove the theorem by *reductio ad absurdum.* We assume that for a certain $\varepsilon_0 > 0$ there is a sequence $\delta_n \to 0+0$ and a sequence $\{ \alpha_n \}$ of numbers α_n corresponding to it which satisfy the condition

$$\frac{1}{\varepsilon (\delta_n)} \delta_n^2 \leqslant \alpha_n \leqslant \alpha (\delta_n), \tag{4.31*}$$

and such that

$$\| X^{\alpha_n} - X^0 \| \geqslant \varepsilon_0 \tag{4.42}$$

for all numbers n.

Since the set $\{ X^\alpha \}$ is bounded, we can isolate a convergent subsequence from the sequence $\{ X^{\alpha_n} \}$ by virtue of the theorem of Bolzano-Weierstrass. In order not to change the designations, we assume that the whole sequence $\{ X^{\alpha_n} \}$ converges to a certain column

$$\bar{X}^0 = \left\| \begin{matrix} \bar{x}_1^0 \\ \cdots \\ \bar{x}_n^0 \end{matrix} \right\|, \text{ i.e. } \| X^{\alpha_n} - \bar{X}^0 \| \to 0 \text{ as } n \to \infty.$$

Let us make sure that

$$\| AX^{\alpha_n} - AX^0 \| \to 0 \text{ as } n \to \infty. \tag{4.43}$$

Indeed, using the triangle inequality, estimates (4.28), (4.29), (4.36) and (4.40) and relations (4.31*) we get

$$\| AX^{\alpha_n} - AX^0 \| \leqslant \| AX^{\alpha_n} - \tilde{A}X^{\alpha_n} \| + \| \tilde{A}X^{\alpha_n} - \tilde{B}_{\tilde{A}} \| + \| \tilde{B}_{\tilde{A}} - AX^0 \|$$

$$\leqslant \delta_n \| X^{\alpha_n} \| + \sqrt{F^{\alpha_n} (X^{\alpha_n}, \tilde{A}, \tilde{B}_{\tilde{A}})} + C\delta_n \leqslant \delta_n (\| X^{\alpha_n} \| + C)$$

$$+ \sqrt{\alpha_n \varepsilon (\delta_n) C + \alpha_n \| X^0 \|^2} \to 0 \text{ as } n \to \infty.$$

It immediately follows from inequality (4.43) that $A\bar{X}^0 = AX^0$, i.e. the limiting element of $\bar{X}^0$ is a solution of system (4.26) satisfying, by virtue of relation (4.41), the inequality $\| \bar{X}^0 \| \leqslant \| X^0 \|$. Since by definition the inverse inequality $\| X^0 \| \leqslant \| \bar{X}^0 \|$ holds true for the normal solution X^0 we have $\| \bar{X}^0 \| = \| X^0 \|$, i.e. $\bar{X}^0 = X^0$, but this contradicts inequality (4.42) which is valid for any number n.

The contradiction we have obtained completes the proof of the theorem.

Chapter 5

LINEAR OPERATORS

This chapter deals with the so-called **linear mappings** of linear and Euclidean spaces, i.e. mappings in which the image of the sum of elements is equal to the sum of their images and the image of the product of an element by a scalar is equal to the product of that scalar by the image of the element. We consider **complex** linear and Euclidean spaces here. The results referring to real spaces will be specified.

5.1. The Concept of a Linear Operator. Its Principal Properties

5.1.1. Definition of a linear operator. Let V and W be linear spaces whose dimensions are n and m respectively. We use the term **operator $\boldsymbol{A}$ from** V into W for the mapping of the form $\boldsymbol{A}: V \to W$, associating every element $\boldsymbol{x}$ of the space V with a certain element $\boldsymbol{y}$ of the space W. In that case we use the designation $\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{x})$ or $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$.

Definition. *The operator $\boldsymbol{A}$ from V into W is said to be linear if for any elements $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ of the space V and any complex number λ the following relations are satisfied:*

1°. $\boldsymbol{A}(\boldsymbol{x}_1 + \boldsymbol{x}_2) = \boldsymbol{A}\boldsymbol{x}_1 + \boldsymbol{A}\boldsymbol{x}_2$ *(the property of additivity of the operator).*

2°. $\boldsymbol{A}(\lambda \boldsymbol{x}) = \lambda \boldsymbol{A}\boldsymbol{x}$ *(the property of homogeneity of the operator).*

Remark 1. *If the space W is a complex plane, then the linear operator $\boldsymbol{A}$ from V into W is called a* ***linear form*** *or* ***linear functional****.*

Remark 2. If the space W coincides with the space V, then the linear operator acting in this case from V into V is also called a **linear transformation of the space** V.

5.1.2. Actions performed on linear operators. The space of linear operators. We shall define, in the set of all linear operators from V into W, the operations of **summation** of such operators and the **multiplication** of an operator by a scalar.

Suppose $\boldsymbol{A}$ and $\boldsymbol{B}$ are two linear operators from V into W. The **sum** of these operators is a linear operator $\boldsymbol{A}+\boldsymbol{B}$ defined by the equation

$$(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{x}=\boldsymbol{A}\boldsymbol{x}+\boldsymbol{B}\boldsymbol{x}. \tag{5.1}$$

The **product** of the linear operator $\boldsymbol{A}$ by the scalar λ is a linear operator $\lambda\boldsymbol{A}$ defined by the equation

$$(\lambda\boldsymbol{A})\boldsymbol{x}=\lambda(\boldsymbol{A}\boldsymbol{x}). \tag{5.2}$$

The **null** operator is an operator, designated as $\boldsymbol{O}$, which maps all the elements of the space V into a null element of the space W. To put it otherwise, the operator $\boldsymbol{O}$ acts according to the rule $\boldsymbol{O}\boldsymbol{x}=\boldsymbol{0}$.

We define the **inverse operator** $-\boldsymbol{A}$ of every operator $\boldsymbol{A}$ by the relation

$$-\boldsymbol{A}=(-1)\boldsymbol{A}.$$

It is easy to verify the validity of the following statement: *The set $L(V, W)$ of all linear operators from V into W, with the operations of summation and multiplication by a scalar indicated above and with a chosen null operator and its inverse forms a linear space.*

5.1.3. The properties of the set $L(V, V)$ of linear operators. Let us discuss, in more detail, linear operators from V into V, i.e. investigate the set $L(V, V)$.

The **identity** (or **unit**) operator is a linear operator $\boldsymbol{I}$ acting according to the rule $\boldsymbol{I}\boldsymbol{x}=\boldsymbol{x}$ (here $\boldsymbol{x}$ is any element of V).

We now introduce the notion of the **product** of linear operators from the set $L(V, V)$.

The **product** of the operators $\boldsymbol{A}$ and $\boldsymbol{B}$ belonging to $L(V, V)$ is an operator $\boldsymbol{AB}$ acting according to the rule

$$(\boldsymbol{AB})\boldsymbol{x}=\boldsymbol{A}(\boldsymbol{B}\boldsymbol{x}). \tag{5.3}$$

Note that in general $\boldsymbol{AB}\neq\boldsymbol{BA}$.

The following properties of linear operators from $L(V, V)$ hold true:

$$\left.\begin{aligned}
&1^\circ.\ \lambda(\boldsymbol{AB})=(\lambda\boldsymbol{A})\boldsymbol{B}.\\
&2^\circ.\ (\boldsymbol{A}+\boldsymbol{B})\boldsymbol{C}=\boldsymbol{AC}+\boldsymbol{BC}.\\
&3^\circ.\ \boldsymbol{A}(\boldsymbol{B}+\boldsymbol{C})=\boldsymbol{AB}+\boldsymbol{AC}.\\
&4^\circ.\ (\boldsymbol{AB})\boldsymbol{C}=\boldsymbol{A}(\boldsymbol{BC}).
\end{aligned}\right\} \tag{5.4}$$

The first property (5.4) follows from the definition of the product of a linear operator by a scalar (see (5.2)) and the definition

of the product of operators (see (5.3)).

Let us substantiate property 2°. In accordance with (5.1), (5.2) and (5.3), we have

$$((\boldsymbol{A}+\boldsymbol{B})\boldsymbol{C})\boldsymbol{x}=(\boldsymbol{A}+\boldsymbol{B})(\boldsymbol{C}\boldsymbol{x})=\boldsymbol{A}(\boldsymbol{C}\boldsymbol{x})+\boldsymbol{B}(\boldsymbol{C}\boldsymbol{x})$$
$$=(\boldsymbol{A}\boldsymbol{C})\boldsymbol{x}+(\boldsymbol{B}\boldsymbol{C})\boldsymbol{x}=(\boldsymbol{A}\boldsymbol{C}+\boldsymbol{B}\boldsymbol{C})\boldsymbol{x}.$$

Comparing the left-hand and the right-hand side of the last relations, we get an equality $(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{C}=\boldsymbol{A}\boldsymbol{C}+\boldsymbol{B}\boldsymbol{C}$.

We have established property 2°.

Property 3° can be proved by a complete analogy.

Property 4° holds true since, in accordance with the definition (see (5.3)), the product of linear operators consists in their succession and, therefore, the linear operators $(\boldsymbol{A}\boldsymbol{B})\boldsymbol{C}$ and $\boldsymbol{A}(\boldsymbol{B}\boldsymbol{C})$ coincide and are consequently identical.

Remark 1. Property 4° makes it possible to define the product $\boldsymbol{A}\boldsymbol{B}\ldots\boldsymbol{C}$ of any finite number of operators from $L(V, V)$ and, in particular, the nth degree of the operator $\boldsymbol{A}$ by the formula

$$\boldsymbol{A}^n=\underbrace{\boldsymbol{A}\boldsymbol{A}\ldots\boldsymbol{A}}_{n \text{ factors}}.$$

The relation $\boldsymbol{A}^{n+m}=\boldsymbol{A}^n\boldsymbol{A}^m$ is obviously true.

We shall need the notion of an **inverse operator** of the given operator $\boldsymbol{A}$ from $L(V, V)$.

Definition 1. *The linear operator $\boldsymbol{B}$ from $L(V, V)$ is said to be an **inverse** of the operator $\boldsymbol{A}$ from $L(V, V)$ if the relation $\boldsymbol{A}\boldsymbol{B}=\boldsymbol{B}\boldsymbol{A}=\boldsymbol{I}$ is satisfied.*

The inverse operator of the operator $\boldsymbol{A}$ is usually designated as $\boldsymbol{A}^{-1}$.

It follows from the definition of the inverse operator $\boldsymbol{A}^{-1}$ that for any $\boldsymbol{x}\in V$ there holds a relation $\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{x}=\boldsymbol{x}$.

Thus, if $\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{x}=\boldsymbol{0}$, then $\boldsymbol{x}=\boldsymbol{0}$, i.e. if the operator $\boldsymbol{A}$ has an inverse, then it follows from the condition $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{0}$ that $\boldsymbol{x}=\boldsymbol{0}$.

We say that the linear operator $\boldsymbol{A}$ is *one-to-one* from V into V if any two distinct elements $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are associated with different elements $\boldsymbol{y}_1=\boldsymbol{A}\boldsymbol{x}_1$ and $\boldsymbol{y}_2=\boldsymbol{A}\boldsymbol{x}_2$.

If the operator $\boldsymbol{A}$ is one-to-one from V into V, then the mapping $\boldsymbol{A}$: $V\to V$ is a mapping of V *onto* V, i.e. *every* element $\boldsymbol{y}\in V$ is an image of a *certain* element $\boldsymbol{x}\in V$:

$$\boldsymbol{y}=\boldsymbol{A}\boldsymbol{x}.$$

To verify this fact, it is evidently sufficient to prove that n linearly independent elements $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ of the space V are mapped by the operator $\boldsymbol{A}$ into n linearly independent elements $\boldsymbol{A}\boldsymbol{x}_1, \boldsymbol{A}\boldsymbol{x}_2, \ldots, \boldsymbol{A}\boldsymbol{x}_n$ of the same space.

Thus, assume that $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ are linearly independent elements of V. If the linear combination $\alpha_1 \boldsymbol{A}\boldsymbol{x}_1 + \alpha_2 \boldsymbol{A}\boldsymbol{x}_2 + \ldots + \alpha_n \boldsymbol{A}\boldsymbol{x}_n$ is a zero element of the space V,

$$\alpha_1 \boldsymbol{A}\boldsymbol{x}_1 + \alpha_2 \boldsymbol{A}\boldsymbol{x}_2 + \ldots + \alpha_n \boldsymbol{A}\boldsymbol{x}_n = \boldsymbol{0},$$

then it follows from the definition of a linear operator (see 5.1.1.) that $\boldsymbol{A}(\alpha_1 \boldsymbol{x}_1 + \alpha_2 \boldsymbol{x}_2 + \ldots + \alpha_n \boldsymbol{x}_n) = \boldsymbol{0}$.

Since the operator $\boldsymbol{A}$ is one-to-one from V into V, it follows from the last relation that $\alpha_1 \boldsymbol{x}_1 + \alpha_2 \boldsymbol{x}_2 + \ldots + \alpha_n \boldsymbol{x}_n = \boldsymbol{0}$. But since the elements $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ are linearly independent, it follows that $\alpha_1 = \alpha_2 = \ldots = \alpha_n = 0$. Consequently the elements $\boldsymbol{A}\boldsymbol{x}_1, \boldsymbol{A}\boldsymbol{x}_2, \ldots$ $\ldots, \boldsymbol{A}\boldsymbol{x}_n$ are also linearly independent.

Note the following **assertion.**

For the linear operator $\boldsymbol{A}$ from $L(V, V)$ to have an inverse, it is necessary and sufficient that this operator be one-to-one from V into V.

Let us ascertain the *necessity* of this condition. Suppose the operator $\boldsymbol{A}$ has an inverse but is not one-to-one from V into V. This means that certain different elements $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, $\boldsymbol{x}_2 - \boldsymbol{x}_1 \neq \boldsymbol{0}$ from V are associated with one and the same element $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_1 = \boldsymbol{A}\boldsymbol{x}_2$. But then $\boldsymbol{A}(\boldsymbol{x}_2 - \boldsymbol{x}_1) = \boldsymbol{0}$ and since A has an inverse, $\boldsymbol{x}_2 - \boldsymbol{x}_1 = \boldsymbol{0}$. But we pointed out above that $\boldsymbol{x}_2 - \boldsymbol{x}_1 \neq \boldsymbol{0}$. The contradiction obtained proves the necessity of the condition given in the assertion.

Let us now prove the *sufficiency* of the condition.

We assume that the operator $\boldsymbol{A}$ is one-to-one from V into V. Then every element $\boldsymbol{y} \in V$ is associated with an element $\boldsymbol{x} \in V$ such that $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$. There is, therefore, an operator $\boldsymbol{A}^{-1}$ possessing the property $\boldsymbol{A}^{-1}\boldsymbol{y} = \boldsymbol{A}^{-1}(\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{x}$. It is easy to verify that $\boldsymbol{A}^{-1}$ is a linear operator. By definition $\boldsymbol{A}^{-1}$ is an inverse of $\boldsymbol{A}$. We have thus proved the sufficiency of the condition given in the assertion.

We introduce now the notions of the **kernel** and the **image** of a linear operator.

Definition 2. *The kernel of a linear operator $\boldsymbol{A}$ is the set of all the elements $\boldsymbol{x}$ of the space V for which $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$.*

The kernel of the linear operator $\boldsymbol{A}$ is symbolized as $\ker \boldsymbol{A}$.

If $\ker \boldsymbol{A} = \boldsymbol{0}$, then the operator $\boldsymbol{A}$ is one-to-one from V into V. Indeed, in this case the condition $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$ yields $\boldsymbol{x} = \boldsymbol{0}$ and this means that different $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are associated with different $\boldsymbol{y}_1 = \boldsymbol{A}\boldsymbol{x}_1$ and $\boldsymbol{y}_2 = \boldsymbol{A}\boldsymbol{x}_2$ (if $\boldsymbol{y}_1$ were equal to $\boldsymbol{y}_2$, then $\boldsymbol{A}(\boldsymbol{x}_2 - \boldsymbol{x}_1)$ would be zero, i.e. $\boldsymbol{x}_1 = \boldsymbol{x}_2$ and the elements $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ would not differ).

Thus, in accordance with the assertion proved above, *the condition* $\ker \boldsymbol{A} = \boldsymbol{0}$ *is necessary and sufficient for the operator $\boldsymbol{A}$ to have an inverse.*

Definition 3. *The* ***image*** *of a linear operator $\boldsymbol{A}$ is the set of all elements $\boldsymbol{y}$ of the space V which can be represented as $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$.*

The image of the linear operator $\boldsymbol{A}$ is symbolized as im $\boldsymbol{A}$*.

Remark 2. Note that if $\ker \boldsymbol{A}=\mathbf{0}$, then $\operatorname{im} \boldsymbol{A}=V$ and vice versa. Therefore, alongside the condition $\ker \boldsymbol{A}=\mathbf{0}$ indicated above, *the condition* $\operatorname{im} \boldsymbol{A}=V$ *is also necessary and sufficient for the operator* $\boldsymbol{A}$ *to have an inverse.*

Remark 3. The kernel $\ker \boldsymbol{A}$ and the image $\operatorname{im} \boldsymbol{A}$ are evidently linear subspaces of the space V. We can, therefore, consider the *dimensions* $\dim(\ker \boldsymbol{A})$ and $\dim(\operatorname{im} \boldsymbol{A})$ of these subspaces.

The following theorem holds true.

Theorem 5.1. *Assume that the dimension* $\dim V$ *of the space* V *is equal to* n *and* $\boldsymbol{A}$ *is a linear operator from* $L(V, V)$. *Then*

$$\dim(\operatorname{im} \boldsymbol{A})+\dim(\ker \boldsymbol{A})=n.$$

Proof. Since $\ker \boldsymbol{A}$ is a subspace of V, we can indicate a subspace V_1 of the space V such that V is a direct sum of V_1 and $\ker \boldsymbol{A}$**. In accordance with Theorem 2.10, $\dim V_1+\dim(\ker \boldsymbol{A})=n$. Therefore, to prove the theorem, it is sufficient to make sure that $\dim V_1=\dim(\operatorname{im} \boldsymbol{A})$.

Suppose $\dim V_1=p$, $\dim(\operatorname{im} \boldsymbol{A})=q$ and $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_q$ is a basis in $\operatorname{im} \boldsymbol{A}$. The linear operator $\boldsymbol{A}$ being one-to-one from V_1 into $\operatorname{im} \boldsymbol{A}$***, every element $\boldsymbol{y}$ from $\operatorname{im} \boldsymbol{A}$ can be put into correspondence with a single element $\boldsymbol{x} \in V_1$ such that $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{y}$. Therefore, elements $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_q$ are defined in V_1 such that $\boldsymbol{A}\boldsymbol{x}_k=\boldsymbol{y}_k$, $k=1, 2, \ldots, q$. The elements $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_q$ are linearly independent since if $\alpha_1\boldsymbol{x}_1+\alpha_2\boldsymbol{x}_2+\ldots+\alpha_q\boldsymbol{x}_q=\mathbf{0}$, then $\boldsymbol{A}(\alpha_1\boldsymbol{x}_1+\alpha_2\boldsymbol{x}_2+\ldots+\alpha_q\boldsymbol{x}_q)=\alpha_1\boldsymbol{y}_1+\alpha_2\boldsymbol{y}_2+\ldots+\alpha_q\boldsymbol{y}_q=\mathbf{0}$, and since the elements $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_q$ are linearly independent, we have $\alpha_1=\alpha_2=\ldots=\alpha_q=0$, i.e. $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_q$ are also linearly independent. Thus there are q linearly independent elements in V_1. *Consequently,* $p \geqslant q$ (recall that $p=\dim V_1$).

---

* The symbol im must be distinguished from Im used to designate the imaginary part of a complex number.

** To verify this, we choose in V a basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ such that the first r vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_r$ form a basis in $\ker \boldsymbol{A}$. Then the linear closure of the vectors $\boldsymbol{e}_{r+1}, \ldots, \boldsymbol{e}_n$ is V_1 (discussed in more detail in Chapter 4).

*** By analogy with linear operators which are one-to-one from V into V, we can introduce the notion of a one-to-one linear operator $\boldsymbol{A}$ from a linear space V into a linear space W. The characteristic feature of these operators is that they put different elements $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ of the space V into correspondence with different elements $\boldsymbol{y}_1=\boldsymbol{A}\boldsymbol{X}_1$ and $\boldsymbol{y}_2=\boldsymbol{A}\boldsymbol{x}_2$ of the space W. This property is inherent in the operator $\boldsymbol{A}$ from the space $\boldsymbol{V}_1$ into the space $\operatorname{im} \boldsymbol{A}$.

Indeed, if $x \in V_1$, $\boldsymbol{x}_2 \in V_1$, $\boldsymbol{x}_2-\boldsymbol{x}_1 \neq 0$, then $\boldsymbol{x}_2-\boldsymbol{x}_1 \in V_1$ and, therefore, $\boldsymbol{A}\boldsymbol{x}_2 \neq \boldsymbol{A}\boldsymbol{x}_1$ ($\boldsymbol{A}\boldsymbol{x}_1 \in \operatorname{im} \boldsymbol{A}$, $\boldsymbol{A}\boldsymbol{x}_2 \in \operatorname{im} \boldsymbol{A}$) since provided we had $\boldsymbol{A}\boldsymbol{x}_2=\boldsymbol{A}\boldsymbol{x}_1$, we would have $\boldsymbol{A}(\boldsymbol{x}_2-x_1)=\mathbf{0}$, i.e. $\boldsymbol{x}_2-\boldsymbol{x}_1 \in \ker \boldsymbol{A}$ and this would contradict the condition of belonging $\boldsymbol{x}_2-\boldsymbol{x}_1 \in V_1$ and the condition $\boldsymbol{x}_2-\boldsymbol{x}_1 \neq 0$ (V_1 and $\ker \boldsymbol{A}$ form a direct sum and, therefore, have only a zero element in common).

Assume that $p > q$. Let us add elements $\boldsymbol{x}_{q+1}, \boldsymbol{x}_{q+2}, \ldots, \boldsymbol{x}_p$ to the linearly independent elements $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_q$ so that $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ form a basis in V_1. Since $p > q$ and $q = \dim(\operatorname{im} \boldsymbol{A})$, the elements $\boldsymbol{A}\boldsymbol{x}_1, \boldsymbol{A}\boldsymbol{x}_2, \ldots, \boldsymbol{A}\boldsymbol{x}_p$, belonging to $\operatorname{im} \boldsymbol{A}$, are linearly dependent and, therefore, there exist numbers $\lambda_1, \lambda_2, \ldots, \lambda_p$, not all equal to zero, such that $\lambda_1 \boldsymbol{A}\boldsymbol{x}_1 + \lambda_2 \boldsymbol{A}\boldsymbol{x}_2 + \ldots + \lambda_p \boldsymbol{A}\boldsymbol{x}_p = \boldsymbol{0}$. Hence $\boldsymbol{A}(\lambda_1 \boldsymbol{x}_1 + \lambda_2 \boldsymbol{x}_2 + \ldots + \lambda_p \boldsymbol{x}_p) = \boldsymbol{0}$. Since $\boldsymbol{A}$ is one-to-one from V_1 into $\operatorname{im} \boldsymbol{A}$, we get from the last equation $\lambda_1 \boldsymbol{x}_1 + \lambda_2 \boldsymbol{x}_2 + \ldots + \lambda_p \boldsymbol{x}_p = \boldsymbol{0}$.

But $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ is a basis in V_1. Therefore, $\lambda_1 = \lambda_2 = \ldots \lambda_p = 0$. It was indicate l somewhat earlier that not all $\lambda_1, \lambda_2, \ldots, \lambda_p$ are equal to zero. Consequently, the assumption $p > q$ leads to a contradiction. Thus $p = q$ and this completes the proof of the theorem.

The following theorem also holds true, which is reciprocal, in a certain sense, of Theorem 5.1.

Theorem 5.2. *Suppose* V_1 *and* V_2 *are two subspaces of the n-dimensional space* V *such that* $\dim V_1 + \dim V_2 = \dim V$. *Then there is a linear operator* $\boldsymbol{A}$ *from* $L(V, V)$ *such that* $V_1 = \operatorname{im} \boldsymbol{A}$ *and* $V_2 = \ker \boldsymbol{A}$.

Proof. Suppose $\dim V_1 = p$ and $\dim V_2 = q$. We choose in the space V a basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ such that the elements $\boldsymbol{e}_{p+1}, \boldsymbol{e}_{p+2} \ldots, \boldsymbol{e}_n$ belong to V_2. Next we take a basis $\boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_p$ in space V_1.

Let us now find the values of the linear operator $\boldsymbol{A}$ on the base vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ of the space V as follows:

$$\boldsymbol{A}\boldsymbol{e}_1 = \boldsymbol{g}_1, \ \boldsymbol{A}\boldsymbol{e}_2 = \boldsymbol{g}_2, \ \ldots, \ \boldsymbol{A}\boldsymbol{e}_p = \boldsymbol{g}_p,$$

$$\boldsymbol{A}\boldsymbol{e}_{p+1} = \boldsymbol{0}, \ \boldsymbol{A}\boldsymbol{e}_{p+2} = \boldsymbol{0}, \ \ldots, \ \boldsymbol{A}\boldsymbol{e}_n = \boldsymbol{0}.$$

Furthermore, if $\boldsymbol{x} = x_1 \boldsymbol{e}_1 + x_2 \boldsymbol{e}_2 + \ldots + x_p \boldsymbol{e}_p + x_{p+1} \boldsymbol{e}_{p+1} + \ldots + x_n \boldsymbol{e}_n$, then $\boldsymbol{A}\boldsymbol{x} = x_1 \boldsymbol{g}_1 + x_2 \boldsymbol{g}_2 + \ldots + x_p \boldsymbol{g}_p$. The operator $\boldsymbol{A}$ is, evidently linear and possesses the required properties. We have proved the theorem.

Let us introduce the notion of the **rank** of the linear operator $\boldsymbol{A}$.

The rank of the linear operator $\boldsymbol{A}$ is the number, symbolized as rank $\boldsymbol{A}$ and equal to rank $\boldsymbol{A} = \dim(\operatorname{im} \boldsymbol{A})$.

Note the following obvious consequence of Theorem 5.1 and of Remark 2 in this subsection.

Corollary of Theorem 5.1. *For the operator* $\boldsymbol{A}$ *from* $L(V, V)$ *to have an inverse* $\boldsymbol{A}^{-1}$, *it is necessary and sufficient that* $\operatorname{rank} \boldsymbol{A} = \dim V = n$.

Suppose $\boldsymbol{A}$ and $\boldsymbol{B}$ are linear operators from $L(V, V)$. The following theorem holds true.

Theorem 5.3. *The following relations hold true*:

$$\operatorname{rank} \boldsymbol{AB} \leqslant \operatorname{rank} \boldsymbol{A}, \ \operatorname{rank} \boldsymbol{AB} \leqslant \operatorname{rank} \boldsymbol{B}.$$

Proof. Let us prove the first of these relations. It is evident that $\operatorname{im} \boldsymbol{AB} \subseteq \operatorname{im} \boldsymbol{A}$*. Therefore, $\dim(\operatorname{im} \boldsymbol{AB}) \leqslant \dim(\operatorname{im} \boldsymbol{A})$, i.e. $\operatorname{rank} \boldsymbol{AB} \leqslant \operatorname{rank} \boldsymbol{A}$.

To prove the second relation, we make use of the following obvious inclusion**: $\ker \boldsymbol{B} \subseteq \ker \boldsymbol{AB}$.

It follows from this inclusion that $\dim(\ker \boldsymbol{B}) \leqslant \dim(\ker \boldsymbol{AB})$. The last inequality, in its turn, yields an inequality $\dim V - \dim(\ker \boldsymbol{AB}) \leqslant \dim V - \dim(\ker \boldsymbol{B})$, and, in accordance with Theorem 5.1, we find from the last inequality that $\dim(\operatorname{im} \boldsymbol{AB}) \leqslant \dim(\operatorname{im} \boldsymbol{B})$, i.e. $\operatorname{rank} \boldsymbol{AB} \leqslant \operatorname{rank} \boldsymbol{B}$. And this is what we had to prove.

Let us prove one more theorem on the ranks of linear operators.

Theorem 5.4. *Assume that* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *are linear operators from* $L(V, V)$ *and* n *is the dimension of* V. Then $\operatorname{rank} \boldsymbol{AB} \geqslant \operatorname{rank} \boldsymbol{A} + \operatorname{rank} \boldsymbol{B} - n$.

Proof. In accordance with Theorem 5.1.

$$\dim(\operatorname{im} \boldsymbol{AB}) + \dim(\ker \boldsymbol{AB}) = n. \tag{5.5}$$

Since $\operatorname{rank} \boldsymbol{AB} = \dim(\operatorname{im} \boldsymbol{AB})$, we get from 5.5

$$\operatorname{rank} \boldsymbol{AB} = n - \dim(\ker \boldsymbol{AB}). \tag{5.6}$$

Since

$$\dim(\ker \boldsymbol{A}) + \dim(\ker \boldsymbol{B}) = 2n - (\operatorname{rank} \boldsymbol{A} + \operatorname{rank} \boldsymbol{B}) \tag{5.7}$$

in accordance with Theorem 5.1, it is sufficient to establish the inequality

$$\dim(\ker \boldsymbol{AB}) \leqslant \dim(\ker \boldsymbol{A}) + \dim(\ker \boldsymbol{B}) \tag{5.8}$$

in order to prove the theorem. Indeed, this inequality and relation (5.6) yield an inequality

$$\operatorname{rank} \boldsymbol{AB} \geqslant n - (\dim(\ker \boldsymbol{A}) + \dim(\ker \boldsymbol{B}))$$

from which, in accordance with (5.7) immediately follows the validity of the assertion of the theorem.

We pass now to the substantiation of inequality (5.8).

---

* Here and in what follows the symbol $\subseteq$ denotes an inclusion, i.e. the notation $\boldsymbol{A} \subseteq \boldsymbol{B}$ means that $\boldsymbol{A}$ is a subset of $\boldsymbol{B}$.

** Since $\boldsymbol{AB}$ and $\boldsymbol{BA}$ are, in general, distinct operators, the inclusion $\operatorname{im} \boldsymbol{AB} \subseteq \operatorname{im} \boldsymbol{B}$ may not occur and therefore, special arguments are required to prove the second relation $\operatorname{rank} \boldsymbol{AB} \leqslant \operatorname{rank} \boldsymbol{B}$.

Assume

$$\dim(\ker \boldsymbol{B}) = q. \tag{5.9}$$

According to Theorem 5.3, $\dim(\ker \boldsymbol{AB}) \geqslant q$. Therefore, the relation

$$\dim(\ker \boldsymbol{AB}) = p + q, \text{ where } p \geqslant 0 \tag{5.10}$$

holds true.

Since $\ker \boldsymbol{B} \subseteq \ker \boldsymbol{AB}$, we can choose a basis $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{p+}$ in the subspace $\ker \boldsymbol{AB}$ such that the elements $\boldsymbol{x}_{p+1}, \ldots, \boldsymbol{x}_{p+q}$ form a basis in $\ker \boldsymbol{B}$. With such a choice of $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{p+q}$ the elements $\boldsymbol{Bx}_1, \boldsymbol{Bx}_2, \ldots, \boldsymbol{Bx}_p$ are linearly independent (if the linear combination $\sum_{k=1}^{p} \lambda_k \boldsymbol{Bx}_k = \boldsymbol{0}$, then $\boldsymbol{B}(\sum_{k=1}^{p} \lambda_k \boldsymbol{x}_k) = \boldsymbol{0}$, i.e. $\sum_{k=1}^{p} \lambda_k \boldsymbol{x}_k \in \ker \boldsymbol{B}$ and this can happen, by virtue of the choice of $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$, only for $\lambda_k = 0$, $k = 1, 2, \ldots, p$). Therefore, the elements $\boldsymbol{Bx}_1, \boldsymbol{Bx}_2, \ldots, \boldsymbol{Bx}_p$ belong to $\ker \boldsymbol{A}$, i.e. $p \leqslant \dim(\ker \boldsymbol{A})$. This inequality and relations (5.9) and (5.10) yield the required inequality (5.8). We have proved the theorem.

Corollary of Theorems 5.3 and 5.4. *If* $\operatorname{rank} \boldsymbol{A} = n$ (n being the dimension of V), *then* $\operatorname{rank} \boldsymbol{AB} = \operatorname{rank} \boldsymbol{BA} = \operatorname{rank} \boldsymbol{B}$.

This corollary follows from the inequalities

$$\operatorname{rank} \boldsymbol{AB} \leqslant \operatorname{rank} \boldsymbol{B} \quad \text{(Theorem 5.3)},$$

$$\operatorname{rank} \boldsymbol{AB} \geqslant \operatorname{rank} \boldsymbol{B} \quad \text{(Theorem 5.4 for rank } \boldsymbol{A} = n).$$

From these inequalities we find that $\operatorname{rank} \boldsymbol{AB} = \operatorname{rank} \boldsymbol{B}$.

The relation $\operatorname{rank} \boldsymbol{BA} = \operatorname{rank} \boldsymbol{B}$ can be proved by analogy.

5.2. Matrix Notation for Linear Operators

5.2.1. The matrices of linear operators in a given basis of the linear space V. We fix the basis $\boldsymbol{e}_1, \boldsymbol{e}_2 \ldots, \boldsymbol{e}_n$ in the linear space V. Suppose $\boldsymbol{x}$ is an arbitrary element of V and

$$\boldsymbol{x} = \sum_{k=1}^{n} x^k \boldsymbol{e}_k \tag{5.11}$$

is the resolution of $\boldsymbol{x}$ into the given components.

Assume that $\boldsymbol{A}$ is a linear operator from $L(V, V)$. Then (5.11) yields

$$\boldsymbol{Ax} = \sum_{k=1}^{n} x^k \boldsymbol{Ae}_k. \tag{5.12}$$

Setting

$$\boldsymbol{A}\boldsymbol{e}_k = \sum_{j=1}^{n} a_k^j \boldsymbol{e}_j, \tag{5.13}$$

we rewrite (5.12) in the following form:

$$\boldsymbol{A}\boldsymbol{x} = \sum_{k=1}^{n} x^k \sum_{j=1}^{n} a_k^j \boldsymbol{e}_j = \sum_{j=1}^{n} \Big(\sum_{k=1}^{n} a_k^j x^k \Big) \boldsymbol{e}_j.$$

Thus, if $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ and the element $\boldsymbol{y}$ has coordinates $y^1, y^2, \ldots, y^n$, then

$$y^j = \sum_{k=1}^{n} a_k^j x^k, \quad j = 1, 2, \ldots, n. \tag{5.14}$$

Let us consider a square matrix A with elements a_k^j:

$$A = (a_k^j).$$

This matrix is known as the **matrix of a linear operator in the given basis** $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$.

Besides the form of notation used for a linear operator indicated above, the matrix notation $\boldsymbol{y} = A\boldsymbol{x}$ is also used for the given basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$, and if in this case $\boldsymbol{x} = (x^1, x^2, \ldots, x^n)$, then $\boldsymbol{y} = (y^1, y^2, \ldots, y^n)$, where y^j, $j = 1, 2, \ldots, n$, are defined by means of relations (5.14) and the elements a_k^j of the matrix A are calculated by formulas (5.13).

Remark 1. If the operator $\boldsymbol{A}$ is zero, then all the elements of the matrix A of that operator are equal to zero in any basis, i.e. A is a null matrix.

Remark 2. If $\boldsymbol{A}$ is an identity operator, i.e. $\boldsymbol{A} = \boldsymbol{I}$, then the matrix of that operator is an identity matrix in any basis. To put it otherwise, in this case $A = I$, where I is an identity matrix.

We have found that for the given basis of the linear space V, every linear operator $\boldsymbol{A}$ from $L(V, V)$ is associated with the matrix A of that operator. The reverse question naturally arises whether with a given basis in V every given matrix A can be put into correspondence with a linear operator $\boldsymbol{A}$ whose matrix is the given matrix. It is also important to find out whether the matrix of the linear operator in the given basis is unique.

The following statement holds true.

Theorem 5.5. *Suppose that a basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ is given in the linear space V and that $A = (a_k^j)$ is a square matrix consisting of n rows and n columns. There is a unique linear operator $\boldsymbol{A}$ whose matrix in the given basis is the matrix A.*

Proof. Let us first prove the existence of the operator $\boldsymbol{A}$. For that purpose we determine the values $\boldsymbol{A}\boldsymbol{e}_k$ of that operator on the base

vectors $\boldsymbol{e}_k$ with the aid of relation (5.13), setting in this relatio a_k^j equal to the corresponding elements of the given matrix A. Th value of the operator $\boldsymbol{A}$ on the arbitrary vector $\boldsymbol{x} \in V$, whose represen tation in terms of the base vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ is given by formul (5.11), can be found by formula (5.12).

The operator we have constructed is, evidently, linear and th matrix A is a matrix of this operator.

The uniqueness of the operator $\boldsymbol{A}$ whose matrix in the basis $\boldsymbol{e}_1, \boldsymbol{e}_2$ $\ldots, \boldsymbol{e}_n$ is the matrix A follows from relations (5.13); these relation make it possible to define uniquely the values of the operator on th base vectors.

Remark 3. Assume that A and B are square matrices of order n, $\boldsymbol{A}$ and $\boldsymbol{B}$ are linear operators in the given basis $\{\boldsymbol{e}_k\}$ of the space V cor responding to them. It follows from the proof of Theorem 5.5. that th matrix $A+\lambda B$, where λ is some number, is associated with a linea operator $\boldsymbol{A}+\lambda\boldsymbol{B}$ (recall that $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{A}+\lambda\boldsymbol{B}$ belong to $L(V, V)$)

Let us prove the following theorem.

Theorem 5.6. *The rank of the linear operator $\boldsymbol{A}$ is equal to the rank of the matrix A of that operator:* rank $\boldsymbol{A}=$ rank A.

Proof. By definition, rank $\boldsymbol{A}=\dim(\operatorname{im}\boldsymbol{A})$ and im $\boldsymbol{A}$ is a linear clo sure of the vectors $\boldsymbol{g}_k$:

$$\boldsymbol{g}_k=\sum_{j=1}^{n} a_k^j \boldsymbol{e}_j \tag{5.15}$$

(see the matrix form of notation for an operator and the definitio of im $\boldsymbol{A}$).

Therefore, rank $\boldsymbol{A}$ is equal to the maximum number of linearl independent vectors $\boldsymbol{g}_k$. Since the vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ are linearl independent, it follows, in accordance with (5.15), that the maximum number of linearly independent vectors $\boldsymbol{g}_k$ coincides with the maxi mum number of linearly independent rows $(a_k^1, a_k^2, \ldots, a_k^n)$ of th matrix A, i.e. with the rank of A. We have proved the theorem

Assume that A and B are arbitrary square matrices consisting o n rows and n columns. Theorems 5.3, 5.4, 5.5 and 5.6 have the fol lowing corollaries.

Corollary 1. The rank AB of the product of A and B satisfies th relations

$$\operatorname{rank} AB \leqslant \operatorname{rank} A, \quad \operatorname{rank} AB \leqslant \operatorname{rank} B,$$

$$\operatorname{rank} AB \geqslant \operatorname{rank} A+\operatorname{rank} B-n.$$

Corollary 2. The inverse operator $\boldsymbol{A}^{-1}$ of the operator $\boldsymbol{A}$ exist only in the case when the rank of the matrix A of the operator $\boldsymbol{A}$ i n ($n=\dim V$). Note that in this case there is also an inverse matri A^{-1} of the matrix A.

5.2.2. Transformation of the matrix of a linear operator in passing to a new basis. Suppose V is a linear space, $\boldsymbol{A}$ is a linear operator from $L(V, V)$, $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ and $\tilde{\boldsymbol{e}}_1, \tilde{\boldsymbol{e}}_2, \ldots, \tilde{\boldsymbol{e}}_n$ are two bases in V and

$$\tilde{\boldsymbol{e}}_k = \sum_{i=1}^{n} u_k^i \boldsymbol{e}_i, \quad k = 1, 2, \ldots, n \tag{5.16}$$

are formulas for passing from the basis $\{\boldsymbol{e}_i\}$ to the basis $\{\tilde{\boldsymbol{e}}_k\}$. Let us denote the matrix (u_k^i) by U:

$$U = (u_k^i). \tag{5.17}$$

Note that rank $U = n$. Suppose

$$A = (a_k^j) \text{ and } \tilde{A} = (\tilde{a}_k^j) \tag{5.18}$$

are the matrices of the operator $\boldsymbol{A}$ in the indicated bases. Let us find the connection between the matrices.

The following statement holds true.

Theorem 5.7. *The matrices A and $\tilde{A}$ of the operator $\boldsymbol{A}$ in the bases $\{\boldsymbol{e}_k\}$ and $\{\tilde{\boldsymbol{e}}_k\}$ respectively are connected by the relation*

$$A = U^{-1} \tilde{A} U,$$

where U^{-1} is an inverse of the matrix U defined by equation* (5.17).

Proof. Directing our attention to the notion of the matrix of a linear operator, we get, according to (5.18),

$$\boldsymbol{A}\boldsymbol{e}_k = \sum_{i=1}^{n} a_k^i \boldsymbol{e}_i, \quad \boldsymbol{A}\tilde{\boldsymbol{e}}_k = \sum_{i=1}^{n} \tilde{a}_k^i \tilde{\boldsymbol{e}}_i. \tag{5.19}$$

The definition of a linear operator, formulas (5.16) and the second formula (5.19) yield relations

$$\boldsymbol{A}\tilde{\boldsymbol{e}}_k = \boldsymbol{A}\left(\sum_{i=1}^{n} u_k^i \boldsymbol{e}_i\right), \quad \boldsymbol{A}\tilde{\boldsymbol{e}}_k = \sum_{i=1}^{n} \tilde{a}_k^i \sum_{j=1}^{n} u_i^j \boldsymbol{e}_j.$$

Therefore, the equality $\sum_{i=1}^{n} u_k^i \boldsymbol{A}\boldsymbol{e}_i = \sum_{j=1}^{n}\left(\sum_{i=1}^{n} \tilde{a}_k^i u_i^j\right)\boldsymbol{e}_j$ holds true.

Substituting the expression $\boldsymbol{A}\boldsymbol{e}_i$ into the left-hand side of this equation in accordance with the first formula (5.19), we find

$$\sum_{j=1}^{n}\left(\sum_{i=1}^{n} u_k^i a_i^j\right)\boldsymbol{e}_j = \sum_{j=1}^{n}\left(\sum_{i=1}^{n} \tilde{a}_k^i u_i^j\right)\boldsymbol{e}_j.$$

Since $\{\boldsymbol{e}_j\}$ is a basis, the last relation yields equations

$$\sum_{i=1}^{n} u_k^i a_i^j = \sum_{i=1}^{n} \tilde{a}_k^i u_i^j, \quad j, k = 1, 2, \ldots, n. \tag{5.20}$$

* Since rank $U = n$, the inverse U^{-1} of the matrix U exists.

If we direct our attention to the matrices A, $\tilde{A}$ and U (see (5.17) and (5.18)), we see that relations (5.20) are equivalent to the following matrix equation: $UA=\tilde{A}U$.

Multiplying both sides of this equation on the left by the matrix U^{-1}, we get the required relation $A=U^{-1}\tilde{A}U$.

And this is what we wished to prove.

Remark 1. Let us consider the formula $A=U^{-1}\tilde{A}U$. Multiplying both sides of the matrix equality on the left by the matrix U and on the right by U^{-1}, we obtain a relation

$$\tilde{A}=UAU^{-1} \tag{5.21}$$

which is another form of connection between the matrices A and $\tilde{A}$ of the linear operator $\boldsymbol{A}$ in different bases.

Remark 2. Suppose A and B are square matrices of order n, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are linear operators in the given basis $\{\boldsymbol{e}_k\}$ corresponding to them. As was pointed out (see Remark 3 in 5.2.1), the matrix $A+\lambda B$ is associated with a linear operator $\boldsymbol{A}+\lambda\boldsymbol{B}$. Let us find the form of the matrix of this operator in the basis $\{\tilde{\boldsymbol{e}}_k\}$. Assume that $\tilde{A}$ and $\tilde{B}$ are the matrices of the operators $\boldsymbol{A}$ and $\boldsymbol{B}$ in the basis $\{\tilde{\boldsymbol{e}}_k\}$. Then, in accordance with (5.21), we have

$$\tilde{A}=UAU^{-1},\ \tilde{B}=UBU^{-1}. \tag{5.22}$$

In accordance with (5.21), the matrix of the linear operator $\boldsymbol{A}+\lambda\boldsymbol{B}$ in the basis $\{\tilde{\boldsymbol{e}}_k\}$ has the form $U(A+\lambda B)U^{-1}$. Using the distributive property of matrix multiplication, we rewrite the last formula as follows (recall that this formula is a matrix of the linear operator $\boldsymbol{A}+\lambda\boldsymbol{B}$ in the basis $\{\tilde{\boldsymbol{e}}_k\}$):

$$UAU^{-1}+\lambda(UBU^{-1}).$$

Considering relations (5.22), we see that the matrix of the operator $\boldsymbol{A}+\lambda\boldsymbol{B}$ in the basis $\{\tilde{\boldsymbol{e}}_k\}$ is written as follows:

$$\tilde{A}+\lambda\tilde{B}.$$

In particular, if B is an identity matrix, $B=I$, then $\tilde{B}=I$ (see Remark 2 in 5.2.1 and Theorem 5.5) and, therefore, the matrix of the linear operator $\boldsymbol{A}+\lambda\boldsymbol{I}$ in the basis $\{\tilde{\boldsymbol{e}}_k\}$ has the form $\tilde{A}=\lambda I$.

Corollary of Theorem 5.7. $\det A=\det\tilde{A}$.

In fact, the determinant of the product of matrices being equal to the product of the determinants of those matrices, it follows from the relation $A=U^{-1}\tilde{A}U$ that

$$\det A=\det U^{-1}\det A\det U. \tag{5.23}$$

Since $\det U^{-1} \det U = 1$, we obtain from relation (5.23) a relation $\det A = \det \tilde{A}$. Thus, the determinant of the matrix of a linear operator does not depend on the choice of a basis. We can, therefore, introduce the notion of the **determinant** $\det \boldsymbol{A}$ **of the linear operator** $\boldsymbol{A}$ setting

$$\det \boldsymbol{A} = \det A, \tag{5.24}$$

where A is the matrix of the linear operator $\boldsymbol{A}$ in any basis.

5.2.3. A characteristic polynomial of a linear operator. Assume that $\boldsymbol{A}$ is a linear operator and $\boldsymbol{I}$ is an identity operator from $L(V, V)$.

Definition. *A polynomial with respect to* λ

$$\det(\boldsymbol{A} - \lambda \boldsymbol{I}) \tag{5.25}$$

is known as a ***characteristic polynomial*** *of the operator* $\boldsymbol{A}$.

Suppose a basis $\{\boldsymbol{e}_k\}$ is given in the basis V and $A = (a_k^j)$ is a matrix of the operator $\boldsymbol{A}$ in that basis. Then, in accordance with (5.24), the characteristic polynomial (5.25) of the operator $\boldsymbol{A}$ is written as follows:

$$\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = \begin{vmatrix} a_1^1 - \lambda & a_1^2 & & a_1^n \\ a_2^1 & a_2^2 - \lambda & & a_2^n \\ \cdot & \cdot & & \cdot \\ a_n^1 & a_n^2 & & a_n^n - \lambda \end{vmatrix}. \tag{5.26}$$

Let us write the characteristic polynomial (5.25) denoting the coefficient in λ^k by d_k:

$$\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = \sum_{k=0}^{n} d_k \lambda^k. \tag{5.27}$$

Remark 1. Since the value of the determinant $\det(\boldsymbol{A} - \lambda \boldsymbol{I})$ does not depend on the choice of a basis, the coefficients d_k of the characteristic polynomial on the right-hand side of (5.27) do not depend on the choice of a basis either. Thus, *the coefficients* d_k *of the characteristic polynomial of the operator* $\boldsymbol{A}$ *are invariants*, i.e. quantities whose values are independent of the choice of a basis.

In particular, the coefficient d_{n-1} evidently equal to $a_1^1 + a_2^2 + \ldots \ldots a_n^n$ is an invariant. This invariant is known as a **trace of the operator** $\boldsymbol{A}$ and is designated as $\operatorname{tr} \boldsymbol{A}$:

$$\operatorname{tr} \boldsymbol{A} = a_1^1 + a_2^2 + \ldots + a_n^n. \tag{5.28}$$

Remark 2. The equation

$$\det(\boldsymbol{A} - \lambda \boldsymbol{L}) = 0 \tag{5.29}$$

is known as a **characteristic equation** of the operator $\boldsymbol{A}$.

5.3. Eigenvalues and Eigenvectors of Linear Operators

Suppose V_1 is a subspace of an n-dimensional linear space V, and $\boldsymbol{A}$ is a linear operator from $L(V, V)$.

Definition 1. *The space V_1 is known as an* ***invariant subspace*** *of the operator $\boldsymbol{A}$ if for every $\boldsymbol{x}$ belonging to V_1 the element $\boldsymbol{Ax}$ also belongs to V_1.*

We can cite ker $\boldsymbol{A}$ and im $\boldsymbol{A}$ as examples of invariant subspaces of the operator $\boldsymbol{A}$.

Definition 2. *The number λ is called an* ***eigenvalue*** *of the operator $\boldsymbol{A}$ if there is a nonzero vector $\boldsymbol{x}$ such that*

$$\boldsymbol{Ax}=\lambda\boldsymbol{x}. \tag{5.30}$$

In this case, the vector $\boldsymbol{x}$ is called an ***eigenvector*** *of the operator $\boldsymbol{A}$ corresponding to the eigenvalue λ.*

The following *statement* holds true.

Theorem 5.8. *For the number λ to be an eigenvalue of the operator $\boldsymbol{A}$, it is necessary and sufficient that this number should be a root of the characteristic equation* (5.29) *of the operator $\boldsymbol{A}$.*

Proof. Suppose λ is an eigenvalue of the operator $\boldsymbol{A}$, and $\boldsymbol{x}$ is an eigenvector corresponding to λ ($\boldsymbol{x}\neq 0$). Let us rewrite relation (5.30) in the following form:

$$(\boldsymbol{A}-\lambda\boldsymbol{I})\,\boldsymbol{x}=0.$$

Since $\boldsymbol{x}$ is a nonzero vector, it follows from the last equation that $\ker(\boldsymbol{A}-\lambda\boldsymbol{I})\neq\boldsymbol{0}$, i.e.

$$\dim(\ker(\boldsymbol{A}-\lambda\boldsymbol{I}))\geqslant 1. \tag{5.31}$$

Since in accordance with Theorem 5.1

$$\dim(\operatorname{im}(\boldsymbol{A}-\lambda\boldsymbol{I}))+\dim(\ker(\boldsymbol{A}-\lambda\boldsymbol{I}))=n,$$

we get from this equation and from inequality (5.31) that

$$\dim(\operatorname{im}(\boldsymbol{A}-\lambda\boldsymbol{I}))\leqslant n-1. \tag{5.32}$$

By definition, dim (im $(\boldsymbol{A}-\lambda\boldsymbol{I})$) is equal to the rank of the operator $\boldsymbol{A}-\lambda\boldsymbol{I}$. Therefore, inequality (5.32) yields

$$\operatorname{rank}(\boldsymbol{A}-\lambda\boldsymbol{I})<n. \tag{5.33}$$

Thus, if λ is an eigenvalue, then the rank of the matrix $A-\lambda I$ of the operator $\boldsymbol{A}-\lambda\boldsymbol{I}$ is smaller than n, i.e. $\det(\boldsymbol{A}-\lambda\boldsymbol{I})=0$ and, consequently, λ is a root of the characteristic equation.

Assume now that λ is a root of the characteristic equation (5.29). Then inequality (5.32) holds true and, consequently, inequality (5.31) also holds true, from which it follows that for the number λ there is a nonzero vector $\boldsymbol{x}$ such that $(\boldsymbol{A}-\lambda\boldsymbol{I})\,\boldsymbol{x}=\boldsymbol{0}$. The last relation is equivalent to relation (5.30). Therefore, λ is an eigenvalue. And this is what we had to prove.

Corollary. *Every linear operator has an eigenvalue.*

Indeed, a characteristic equation always possesses a root (by virtue of the fundamental theorem of algebra).

The following theorem holds true.

Theorem 5.9. *For the matrix A of the linear operator $\boldsymbol{A}$ in the given basis $\{\boldsymbol{e}_k\}$ to be diagonal*, it is necessary and sufficient that the base vectors $\boldsymbol{e}_k$ be eigenvectors of that operator.*

Proof. Suppose the base vectors $\boldsymbol{e}_k$ are eigenvectors of the operator $\boldsymbol{A}$. Then

$$\boldsymbol{A}\boldsymbol{e}_k=\lambda_k\boldsymbol{e}_k, \tag{5.34}$$

and, therefore, the matrix A of the operator $\boldsymbol{A}$ has the form (see relations (5.13), and the concept of the matrix of a linear operator)

$$A=\begin{pmatrix} \lambda_1 & 0 & & 0 \\ 0 & \lambda_2 & & 0 \\ \cdot & \cdot & & \cdot \\ 0 & 0 & & \lambda_n \end{pmatrix} \tag{5.35}$$

i.e. is diagonal.

Suppose the matrix A of the linear operator $\boldsymbol{A}$ in the given basis $\{\boldsymbol{e}_k\}$ is diagonal, i.e. has the form (5.35). Then relations (5.13) assume the form (5.34), and this means that $\boldsymbol{e}_k$ are eigenvectors of the operator $\boldsymbol{A}$. We have proved the theorem.

We shall prove one more property of eigenvectors.

Theorem 5.10. *Suppose the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ of the operator $\boldsymbol{A}$ are distinct. Then the eigenvectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p$ corresponding to them are linearly independent.*

Proof. We reason by induction. Since $\boldsymbol{e}_1$ is a nonzero vector, the statement holds true for one vector ($p=1$), one nonzero vector being linearly independent. Suppose we have proved the theorem for m vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_m$. We add a vector $\boldsymbol{e}_{m+1}$ to these vectors and assume that there holds an equality

$$\sum_{k=1}^{m+1} \alpha_k\boldsymbol{e}_k=0. \tag{5.36}$$

Then, using the properties of linear operator, we get

$$\sum_{k=1}^{m+1} \alpha_k\boldsymbol{A}\boldsymbol{e}_k=0. \tag{5.37}$$

Since $\boldsymbol{e}_k$ are eigenvectors, we have $\boldsymbol{A}\boldsymbol{e}_k=\lambda_k\boldsymbol{e}_k$ and, therefore, equa-

* Recall that a matrix is said to be diagonal if all its elements lying off the principal diagonal are equal to zero.

tion (5.37) can be rewritten as follows:

$$\sum_{k=1}^{m+1} \alpha_k \lambda_k \boldsymbol{e}_k = 0. \tag{5.38}$$

In accordance with (5.36), $\sum_{k=1}^{m+1} \lambda_{m+1} \alpha_k \boldsymbol{e}_k = 0$. Subtracting this equality from equality (5.38), we find that

$$\sum_{k=1}^{m} (\lambda_k - \lambda_{m+1}) \alpha_k \boldsymbol{e}_k = 0. \tag{5.39}$$

All λ_k are distinct by the hypothesis, i.e. $\lambda_k - \lambda_{m+1} \neq 0$. Therefore it follows from (5.39) and from the assumption of a linear independence of the vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_m$ that $\alpha_1 = \alpha_2 = \ldots \alpha_m = 0$. From this equality and from (5.36), and also from the condition that $\boldsymbol{e}_{m+1}$ is an eigenvector ($\boldsymbol{e}_{m+1} \neq \boldsymbol{0}$) it follows that $\alpha_{m+1} = 0$. Thus, we get from equation (5.36) $\alpha_1 = \alpha_2 = \ldots = \alpha_{m+1} = 0$. This means that the vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_{m+1}$ are linearly independent. We have carried out induction and completed the proof of the theorem.

Corollary. *If the characteristic polynomial of the operator* $\boldsymbol{A}$ *has n distinct roots, then in some basis the matrix of the operator* $\boldsymbol{A}$ *has a diagonal form.*

Indeed, in the case being considered, in accordance with the theorem we have just proved, the eigenvectors are linearly independent and can, therefore, be chosen as base vectors. But then, by Theorem 5.9, the matrix of the operator $\boldsymbol{A}$ is diagonal in that basis.

5.4. Linear and Sesquilinear Forms in a Euclidean Space

5.4.1. A special representation of a linear form in a Euclidean space. Suppose V is a Euclidean space, and C is a complex plane (one-dimensional complex linear space).

We introduced the notion of a **linear form,** i.e. a linear operator from V to C in 5.1.1. Here we shall get a special representation of an arbitrary linear form f from $L(V, C)$.

Lemma. *Assume that f is a linear form from* $L(V, C)$. *Then there is a unique element* $\boldsymbol{h}$ *from V such that*

$$f(\boldsymbol{x}) = (\boldsymbol{x}, \boldsymbol{h}). \tag{5.40}$$

Proof. To prove the existence of the element $\boldsymbol{h}$, we choose an orthonormal basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ in V.

Let us consider the element $\boldsymbol{h}$ whose coordinates h^k in the chosen

basis are defined by the relations*

$$h^k = \overline{f(e_k)}. \tag{5.41}$$

Thus, $\boldsymbol{h} = \sum_{k=1}^{n} h^k \boldsymbol{e}_k$.

Assume that $\boldsymbol{x} = \sum_{k=1}^{n} x^k \boldsymbol{e}_k$ is an arbitrary element of the space V. Using the properties of the linear form f and equation (5.41), we obtain

$$f(\boldsymbol{x}) = \sum_{k=1}^{n} x^k f(\boldsymbol{e}_k) = \sum_{k=1}^{n} x^k \bar{h}^k. \tag{5.42}$$

Since in the orthonormal basis $\{\boldsymbol{e}_k\}$ the scalar product $(\boldsymbol{x}, \boldsymbol{h})$ of the vectors $\boldsymbol{x} = \sum_{k=1}^{n} x^k \boldsymbol{e}_k$ and $\boldsymbol{h} = \sum_{k=1}^{n} h^k \boldsymbol{e}_k$ is equal to $\sum_{k=1}^{n} x^k \bar{h}^k$, we get $f(\boldsymbol{x}) = (\boldsymbol{x}, \boldsymbol{h})$ from (5.42).

We have proved the existence of the vector $\boldsymbol{h}$.

Let us now prove the uniqueness of this vector. Assume that $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ are two vectors such that with their aid the form $f(\boldsymbol{x})$ can be represented as (5.40). It is evident that for any $\boldsymbol{x}$ there holds a relation $(\boldsymbol{x}, \boldsymbol{h}_1) = (\boldsymbol{x}, \boldsymbol{h}_2)$ which yields an equality $(\boldsymbol{x}, \boldsymbol{h}_2 - \boldsymbol{h}_1) = 0$. Setting $\boldsymbol{x} = \boldsymbol{h}_2 - \boldsymbol{h}_1$ in this equality and using the definition of the norm of an element in a Euclidean space, we find that $\| \boldsymbol{h}_2 - \boldsymbol{h}_1 \| = 0$. Thus, $\boldsymbol{h}_2 = \boldsymbol{h}_1$ and this proves the lemma.

Remark. The lemma is evidently also valid when V is a real Euclidean space and $f \in L(V, R)$, where R is a real straight line.

5.4.2. Sesquilinear forms in a Euclidean space. Special representation of such forms. We introduce the concept of a **sesquilinear form** in a linear space.

Definition. *The number function $B(\boldsymbol{x}, \boldsymbol{y})$ whose arguments are various vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ of the linear space L is known as a sesquilinear form if the relations*

$$\left.\begin{aligned} B(\boldsymbol{x}+\boldsymbol{y}, \boldsymbol{z}) &= B(\boldsymbol{x}, \boldsymbol{z}) + B(\boldsymbol{y}, \boldsymbol{z}), \\ B(\boldsymbol{x}, \boldsymbol{y}+\boldsymbol{z}) &= B(\boldsymbol{x}, \boldsymbol{y}) + B(\boldsymbol{x}, \boldsymbol{z}), \\ B(\lambda\boldsymbol{x}, \boldsymbol{y}) &= \lambda B(\boldsymbol{x}, \boldsymbol{y}), \\ B(\boldsymbol{x}, \lambda\boldsymbol{y}) &= \bar{\lambda} B(\boldsymbol{x}, \boldsymbol{y}) \end{aligned}\right\} \tag{5.43}$$

hold for any vectors $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$ from L and any complex number λ.

In other words, the sesquilinear form $B(\boldsymbol{x}, \boldsymbol{y})$ is a number function of two vector arguments $\boldsymbol{x}$, $\boldsymbol{y}$ defined on various vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ of

* The bar over $f(\boldsymbol{e}_k)$ signifies that we take a complex conjugate value of this expression.

the linear space L, which is linear with respect to the first argument $\boldsymbol{x}$ and antilinear with respect to the second argument $\boldsymbol{y}$.

Remark 1. If the linear space L is real, then the sesquilinear forms pass into the so-called bilinear forms, i.e. forms which are linear with respect to every argument (since λ is real, the fourth relation (5.43) characterizes the linearity with respect to the second argument as well). Bilinear forms will be discussed in Chapter 7.

Let us consider a sesquilinear form given in a Euclidean space V. The following theorem on a special representation of such a form holds true.

Theorem 5.11. *Suppose $B(\boldsymbol{x}, \boldsymbol{y})$ is a sesquilinear form in the Euclidean space V. Then there is a unique linear operator $\boldsymbol{A}$ from $L(V, V)$ such that*

$$B(\boldsymbol{x}, \boldsymbol{y})=(\boldsymbol{x}, \boldsymbol{A}\boldsymbol{y}). \tag{5.44}$$

Proof. Assume that $\boldsymbol{y}$ is any fixed element of the space V. Then $B(\boldsymbol{x}, \boldsymbol{y})$ is a linear form of the argument $\boldsymbol{x}$. Therefore, in accordance with the lemma given in 5.4.1, there is a well defined element $\boldsymbol{h}$ of the space V such that

$$B(\boldsymbol{x}, \boldsymbol{y})=(\boldsymbol{x}, \boldsymbol{h}). \tag{5.45}$$

Thus, equation (5.45) puts every $\boldsymbol{y}$ from V into correspondence with a unique element $\boldsymbol{h}$ from V. We have thus defined the operator $\boldsymbol{A}$ such that $\boldsymbol{h}=\boldsymbol{A}\boldsymbol{y}$. The linearity of this operator follows directly from properties (5.43) of the sesquilinear form and from the properties of a scalar product.

Let us prove that the operator $\boldsymbol{A}$ is unique.

Suppose $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are two operators such that with their aid the form $B(\boldsymbol{x}, \boldsymbol{y})$ can be represented in the form (5.44). It is evident that the relation $(\boldsymbol{x}, \boldsymbol{A}_1\boldsymbol{y})=(\boldsymbol{x}, \boldsymbol{A}_2\boldsymbol{y})$ holds for any $\boldsymbol{x}$ and $\boldsymbol{y}$, and it follows from the indicated relation that $(\boldsymbol{x}, \boldsymbol{A}_2\boldsymbol{y}-\boldsymbol{A}_1\boldsymbol{y})=0$. Setting $\boldsymbol{x}=\boldsymbol{A}_2\boldsymbol{y}-\boldsymbol{A}_1\boldsymbol{y}$ in this equality and using the definition of the norm of an element we find that $\| \boldsymbol{A}_2\boldsymbol{y}-\boldsymbol{A}_1\boldsymbol{y} \|=0$.

Thus the equality $\boldsymbol{A}_2\boldsymbol{y}=\boldsymbol{A}_1\boldsymbol{y}$ is valid for any $\boldsymbol{y}$ from V, i.e. $\boldsymbol{A}_2=\boldsymbol{A}_1$. We have proved the theorem.

Corollary. *Suppose $B(\boldsymbol{x}, \boldsymbol{y})$ is a sesquilinear form in the Euclidean space V. Then there is a unique linear operator $\boldsymbol{A}$ from $L(V, V)$ such that*

$$B(\boldsymbol{x}, \boldsymbol{y})=(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}). \tag{5.46}$$

The validity of this corollary follows from the following arguments. First, the form $B_1(\boldsymbol{x}, \boldsymbol{y})=\overline{B(\boldsymbol{x}, \boldsymbol{y})}$ is sesquilinear (this follows from the fact that $B(\boldsymbol{x}, \boldsymbol{y})$ is a sesquilinear form and from the definition of such a form). Furthermore, in accordance with Theorem 5.11 we get, for $B_1(\boldsymbol{y}, \boldsymbol{x})$ a representation in the form

$$B_1(\boldsymbol{y}, \boldsymbol{x}) = (\boldsymbol{y}, \boldsymbol{A}\boldsymbol{x}). \tag{5.47}$$

Since the conjugate value of $\overline{B_1(\boldsymbol{x}, \boldsymbol{y})}$ is equal to $B_1(\boldsymbol{x}, \boldsymbol{y})$, we can take the conjugate values of the left-hand and right-hand sides of (5.47) and, bearing in mind that $B_1(\boldsymbol{y}, \boldsymbol{x}) = \overline{B(\boldsymbol{x}, \boldsymbol{y})}$ we obtain

$$B(\boldsymbol{x}, \boldsymbol{y}) = \overline{(\boldsymbol{y}, \boldsymbol{A}\boldsymbol{x})}. \tag{5.48}$$

But $\overline{(\boldsymbol{y}, \boldsymbol{A}\boldsymbol{x})} = (\boldsymbol{A}\boldsymbol{x}, \boldsymbol{y})$ (see 4.3.1.). Therefore, (5.48) yields equation (5.46). We have proved the corollary.

Remark 2. Theorem 5.11 and the corollary of this theorem are also valid for the case when the Euclidean space is real. In that case, in the formulation of the theorem and the corollary the term "sesquilinear form" must be replaced by the term "bilinear form". See also Remark 1.

Let us introduce the concept of a matrix of a sesquilinear form in the given basis $\{\boldsymbol{e}_k\}$.

Assume that $\boldsymbol{x}$ and $\boldsymbol{y}$ belong to V, and $\boldsymbol{x} = \sum_{j=1}^{n} x^j \boldsymbol{e}_j, \boldsymbol{y} = \sum_{k=1}^{n} y^k \boldsymbol{e}_k$ are the resolution of $\boldsymbol{x}$ and $\boldsymbol{y}$ in terms of the base vectors $\{\boldsymbol{e}_k\}$. The definition of a sesquilinear form yields relations

$$B(\boldsymbol{x}, \boldsymbol{y}) = B\left(\sum_{j=1}^{n} x^j \boldsymbol{e}_j, \sum_{k=1}^{n} y^k \boldsymbol{e}_k\right) = \sum_{j=1}^{n} \sum_{k=1}^{n} x^j \bar{y}^k B(\boldsymbol{e}_j, \boldsymbol{e}_k). \tag{5.49}$$

Setting

$$b_{jk} = B(\boldsymbol{e}_j, \boldsymbol{e}_k), \tag{5.50}$$

we write expression (5.49) for $B(\boldsymbol{x}, \boldsymbol{y})$ in the form

$$B(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j,k=1}^{n} b_{jk} x^j \bar{y}^k.$$

The matrix $B = (b_{jk})$ is said to be the ***matrix of the sesquilinear form*** *$B(\boldsymbol{x}, \boldsymbol{y})$ in the basis $\{\boldsymbol{e}_k\}$.*

The following statement holds true.

Suppose that the sesquilinear form $B(\boldsymbol{x}, \boldsymbol{y})$ is represented in the form (5.46)

$$B(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}). \tag{5.46}$$

Suppose, furthermore, that the elements of the matrix A of the operator $\boldsymbol{A}$ in the given orthonormal basis are equal to a_j^k. Then in this basis $b_{jk} = a_j^k$.

To prove this statement, let us consider expression (5.50) for the coefficients b_{jk} in the sesquilinear form. We transform the right-hand side (5.50) with the aid of (5.46). In accordance with (5.13), we get

$$b_{jk}=B(\boldsymbol{e}_j,\ \boldsymbol{e}_k)=(\boldsymbol{A}\boldsymbol{e}_j,\ \boldsymbol{e}_k)=\left(\sum_{q=1}^{n} a_j^q \boldsymbol{e}_q, \boldsymbol{e}_k\right)=\sum_{q=1}^{n} a_j^q(\boldsymbol{e}_q,\ \boldsymbol{e}_k).$$

The basis $\{\boldsymbol{e}_k\}$ being orthonormal, $(\boldsymbol{e}_q, \boldsymbol{e}_k)=0$ if $q\neq k$ and $(\boldsymbol{e}_k, \boldsymbol{e}_k)=1$. Therefore, from all the terms of the last sum, only that term is nonzero which results at $q=k$. Thus, $b_{jk}=a_j^k$. We have thus proved the statement.

Remark 3. If the sesquilinear form is represented as $B(\boldsymbol{x},\ \boldsymbol{y})=(\boldsymbol{x},\ \boldsymbol{A}\boldsymbol{y})$ and the elements of the matrix A of the operator $\boldsymbol{A}$ in the given orthonormal basis are equal to a_j^k, then $b_{jk}=\bar{a}_j^k$ in that basis.

5.5. Linear Self-Adjoint Operators in a Euclidean Space

5.5.1. The concept of an adjoint operator. We shall consider linear operators in a finite-dimensional Euclidean space V.

Definition 1. *The operator $\boldsymbol{A}^*$ from $L(V, V)$ is said to be an* ***adjoint*** *or a* ***conjugate*** *operator of the linear operator $\boldsymbol{A}$ if the relation*

$$(\boldsymbol{A}\boldsymbol{x},\ \boldsymbol{y})=(\boldsymbol{x},\ \boldsymbol{A}^*\boldsymbol{y}) \tag{5.51}$$

holds for any $\boldsymbol{x}$ and $\boldsymbol{y}$ from V.

It is easy to verify that the operator $\boldsymbol{A}^*$, which is an adjoint of the linear operator $\boldsymbol{A}$, is itself a linear operator. This follows from the obvious relation

$$\begin{aligned}(\boldsymbol{A}\boldsymbol{x},\ \alpha\boldsymbol{y}_1+\beta\boldsymbol{y}_2)&=\bar{\alpha}(\boldsymbol{A}\boldsymbol{x},\ \boldsymbol{y}_1)+\bar{\beta}(\boldsymbol{A}\boldsymbol{x},\ \boldsymbol{y}_2)\\&=\bar{\alpha}(\boldsymbol{x},\ \boldsymbol{A}^*\boldsymbol{y}_1)+\bar{\beta}(\boldsymbol{x},\ \boldsymbol{A}^*\boldsymbol{y}_2)=(\boldsymbol{x},\ \boldsymbol{A}^*(\alpha\boldsymbol{y}_1+\beta\boldsymbol{y}_2)),\end{aligned}$$

which is valid for any element $\boldsymbol{x}$, $\boldsymbol{y}_1$, $\boldsymbol{y}_2$ and any complex numbers α and β.

Let us prove the following theorem.

Theorem 5.12. *Every linear operator $\boldsymbol{A}$ has a single adjoint.*

Proof. The scalar product $(\boldsymbol{A}\boldsymbol{x},\ \boldsymbol{y})$ is evidently a sesquilinear form (see 4.3.1 and the definition of a sesquilinear form).

By Theorem 5.11, there is a unique linear operator $\boldsymbol{A}^*$ such that this form can be represented as $(\boldsymbol{x},\ \boldsymbol{A}^*\boldsymbol{y})$. Thus, $(\boldsymbol{A}\boldsymbol{x},\ \boldsymbol{y})=(\boldsymbol{x},\ \boldsymbol{A}^*\boldsymbol{y})$.

Consequently, the operator $\boldsymbol{A}^*$ is an adjoint of the operator $\boldsymbol{A}$. The uniqueness of $\boldsymbol{A}^*$ follows from the uniqueness of the representation of a sesquilinear operator in the form (5.44). And this is what we wished to prove.

In what follows the symbol $\boldsymbol{A}^*$ designates an operator which is an adjoint of $\boldsymbol{A}$.

Note the following properties of adjoint operators:

1°. $I^*=I$.

2°. $(A+B)^*=A^*+B^*$.

3°. $(\lambda A)^*=\bar{\lambda}A^*$.

4°. $(A^*)^*=A$.

5°. $(AB)^*=B^*A^*$.

It is easy to prove properties 1°-4° and we invite the reader to do so. Here is the proof of property 5°.

In accordance with the definition of a product of operators, the relation $(AB)x=A(Bx)$ holds true. With the aid of this equality and the definition of an adjoint operator, we get the following chain of relations:

$$((AB)x,\ y)=(A(Bx),y)=(Bx,\ A^*y)$$
$$=(x,\ B^*(A^*y))=(x,(B^*A^*)y).$$

Thus, $((AB)x,\ y)=(x,\ (B^*A^*)y)$. In other words, the operator B^*A^* is an adjoint of the operator AB. We have proved the validity of property 5°.

Remark. The notion of an adjoint operator for a real space can be introduced by analogy. The inferences made here and the properties of adjoint operators are also valid in this case (here property 3° can be formulated as $(\lambda A)^*=\lambda A^*$).

5.5.2. Self-adjoint operators. Basic properties.

Definition 2. *The linear operator A from $L(V,\ V)$ is* ***self-adjoint*** *if the relation*

$$A^*=A$$

holds true. A self-adjoint operator in a real space can be defined by analogy.

An identity operator I can serve as a simple example of a self-adjoint operator (see property 1° of adjoint operators in 5.5.1).

Self-adjoint operators help in obtaining a special representation of arbitrary linear operators. Namely, the following statement holds true.

Theorem 5.13. *Let A be a linear operator acting in a complex Euclidean space V. Then the representation $A=A_R+iA_I$ holds true, where A_R and A_I are self-adjoint operators called the real and the imaginary part of the operator A respectively.*

Proof. In accordance with properties 2°, 3° and 4° of adjoint operators (see 5.5.1), the operators $A_R=(A+A^*)/2$ and $A_I=(A-A^*)/2i$ are self-adjoint.

Evidently, $A=A_R+iA_I$. We have proved the theorem.

In the following theorem we elucidate the conditions under which the products of self-adjoint operators are self-adjoint. We say that the operators A and B *commute* if $AB=BA$.

Theorem 5.14. *For the product* $\boldsymbol{AB}$ *of the self-adjoint operators* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *to be a self-adjoint operator, it is necessary and sufficient that they should commute.*

Proof. $\boldsymbol{A}$ and $\boldsymbol{B}$ being self-adjoint operators, the relations

$$(\boldsymbol{AB})^* = \boldsymbol{B}^*\boldsymbol{A}^* = \boldsymbol{BA} \tag{5.52}$$

are valid in accordance with property 5° of adjoint operators (see 5.5.1).

Consequently, if $\boldsymbol{AB} = \boldsymbol{BA}$, then $(\boldsymbol{AB})^* = \boldsymbol{AB}$, i.e. the operator $\boldsymbol{AB}$ is self-adjoint. Now if $\boldsymbol{AB}$ is a self-adjoint operator, then $\boldsymbol{AB} = (\boldsymbol{AB})^*$, and then, by virtue of (5.52), $\boldsymbol{AB} = \boldsymbol{BA}$. And this is what we set out to prove.

In the theorems that follow we establish a number of significant properties of self-adjoint operators.

Theorem 5.15. *If the operator* $\boldsymbol{A}$ *is self-adjoint, then for any* $\boldsymbol{x} \in V$ *the scalar product* $(\boldsymbol{Ax}, \boldsymbol{x})$ *is a real number.*

Proof. The validity of the statement of the theorem follows from the property of a scalar product in a complex Euclidean space $(\boldsymbol{Ax}, \boldsymbol{x}) = \overline{(\boldsymbol{x}, \boldsymbol{Ax})}$ and the definition of a self-adjoint operator $(\boldsymbol{Ax}, \boldsymbol{x}) = (\boldsymbol{x}, \boldsymbol{Ax})^*$.

Theorem 5.16. *The eigenvalues of a self-adjoint operator are real.*

Proof. Let λ be an eigenvalue of the self-adjoint operator $\boldsymbol{A}$. By definition of an eigenvalue of the operator $\boldsymbol{A}$ (see Definition 2 in 5.3), there is a nonzero vector $\boldsymbol{x}$ such that $\boldsymbol{Ax} = \lambda\boldsymbol{x}$. It follows from this relation that the real (by virtue of Theorem 5.15) scalar product $(\boldsymbol{Ax}, \boldsymbol{x})$ can be represented as

$$(\boldsymbol{Ax}, \boldsymbol{x}) = \lambda(\boldsymbol{x}, \boldsymbol{x}) = \lambda \| \boldsymbol{x} \|^{**}.$$

Since $\| \boldsymbol{x} \|$ and $(\boldsymbol{Ax}, \boldsymbol{x})$ are real, λ is evidently also real. We have proved the theorem.

In the following theorem we elucidate the property of orthogonality of the eigenvectors of a self-adjoint operator.

Theorem 5.17. *If* $\boldsymbol{A}$ *is a self-adjoint operator, then the eigenvectors corresponding to various eigenvalues of that operator are orthogonal.*

Proof. Let λ_1 and λ_2 be various eigenvalues ($\lambda_1 \neq \lambda_2$) of the self-adjoint operator $\boldsymbol{A}$, and $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ be the eigenvectors corresponding to them. Then there hold relations $\boldsymbol{Ax}_1 = \lambda_1\boldsymbol{x}_1$, $\boldsymbol{Ax}_2 = \lambda_2\boldsymbol{x}_2$. Therefore, the scalar products $(\boldsymbol{Ax}_1, \boldsymbol{x}_2)$ and $(\boldsymbol{x}_1, \boldsymbol{Ax}_2)$ are respectively equal to

---

* Recall that if a complex number is equal to its conjugate number, then that number is real.

** Recall that the symbol $\| \boldsymbol{x} \|$ is used to designate the norm of the element $\boldsymbol{x}$.

the expressions

$$(\boldsymbol{A}\boldsymbol{x}_1,\ \boldsymbol{x}_2)=\lambda_1(\boldsymbol{x}_1,\ \boldsymbol{x}_2),\ (\boldsymbol{x}_1,\ \boldsymbol{A}\boldsymbol{x}_2)=\lambda_2\,(\boldsymbol{x}_1,\ \boldsymbol{x}_2)^*.$$

The operator $\boldsymbol{A}$ being self-adjoint, the scalar products $(\boldsymbol{A}\boldsymbol{x}_1,\ \boldsymbol{x}_2)$ and $(\boldsymbol{x}_1,\ \boldsymbol{A}\boldsymbol{x}_2)$ are equal and, therefore, by means of subtraction we get from the last relations an equation

$$(\lambda_2-\lambda_1)\,(\boldsymbol{x}_1,\ \boldsymbol{x}_2)=0.$$

Since $\lambda_2\neq\lambda_1$, it follows from the last equation that the scalar product $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ is equal to zero, i.e. that the eigenvectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are orthogonal. And this is what we wished to prove.

5.5.3. The norm of a linear operator. Let $\boldsymbol{A}$ be a linear operator mapping the Euclidean space V into the same space. We introduce the concept of the norm of the operator $\boldsymbol{A}$.

Definition 3. *The norm* $\|\boldsymbol{A}\|$ *of the linear operator* $\boldsymbol{A}$ *is a number defined by the relation***

$$\|\boldsymbol{A}\|=\sup_{\|x\|=1}\|\boldsymbol{A}\boldsymbol{x}\|. \tag{5.53}$$

The definition of the norm of a linear operator yields the following obvious inequality:

$$\|\boldsymbol{A}\boldsymbol{x}\|\leqslant\|\boldsymbol{A}\|\,\|\boldsymbol{x}\| \tag{5.54}$$

(to prove it, it is sufficient to use the relation $\boldsymbol{A}\boldsymbol{x}=\left(\boldsymbol{A}\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\right)\|\boldsymbol{x}\|$). It follows from relations (5.54) that if $\|\boldsymbol{A}\|=0$, then the operator $\boldsymbol{A}$ is *zero*.

We can define the norm of the self-adjoint operator $\boldsymbol{A}$ in another way. Namely, the following *statement* holds true.

If $\boldsymbol{A}$ *is a self-adjoint operator, then the norm* $\|\boldsymbol{A}\|$ *of the operator* $\boldsymbol{A}$ *introduced above is* $\sup\limits_{\|x\|=1}|(\boldsymbol{A}\boldsymbol{x},\ \boldsymbol{x})|$:

$$\sup_{\|x\|=1}|\,(\boldsymbol{A}\boldsymbol{x},\ \boldsymbol{x})\,|=\|\boldsymbol{A}\|. \tag{5.55}$$

Proof. The Cauchy-Bunyakovsky inequality (see 4.3.2)

$$|\,(\boldsymbol{A}\boldsymbol{x},\ \boldsymbol{x})\,|\leqslant\|\boldsymbol{A}\boldsymbol{x}\|\,\|\boldsymbol{x}\|$$

is valid for any $\boldsymbol{x}$ from V. From this inequality and inequality (5.54) we get an inequality $|\,(\boldsymbol{A}\boldsymbol{x},\ \boldsymbol{x})\,|\leqslant\|\boldsymbol{A}\|\,\|\boldsymbol{x}\|^2$. Therefore, the number

$$\mu=\sup_{\|x\|=1}|\,(\boldsymbol{A}\boldsymbol{x},\ \boldsymbol{x})\,| \tag{5.56}$$

---

* Since the eigenvalues of a self-adjoint operator are real, we have $(\boldsymbol{x}_1,\ \boldsymbol{A}\boldsymbol{x}_2)=\bar{\lambda}_2(\boldsymbol{x}_1,\ \boldsymbol{x}_2)=\lambda_2(\boldsymbol{x}_1,\ \boldsymbol{x}_2)$.

** Recall that $\|\boldsymbol{A}\boldsymbol{x}\|=\sqrt{(\boldsymbol{A}\boldsymbol{x},\ \boldsymbol{A}\boldsymbol{x})}$. Hence it follows that $\|\boldsymbol{A}\boldsymbol{x}\|$ is a continuous function of $\boldsymbol{x}$ which attains its finite greatest value on the closed set $\|\boldsymbol{x}\|=1$.

satisfies the relation

$$\mu \leqslant \|A\|. \tag{5.57}$$

Note that the equation $(Az, z)=\left(A\frac{z}{\|z\|}, \frac{z}{\|z\|}\right)\|z\|^2$, $z\neq 0$, and the definition of the number μ (see (5.56)) yields an inequality

$$|(Az, z)| \leqslant \mu \|z\|^2. \tag{5.58}$$

Let us draw attention to the obvious identity

$$4\operatorname{Re}(Ax, y)=(A(x+y), x+y)-(A(x-y), x-y)$$

(the symbol $\operatorname{Re}(Ax, y)$ in this identity is used for the real part of the complex number (Ax, y); the identity itself immediately follows from the properties of a scalar product, see 4.3.1.). Taking the absolute values of the expressions on the left-hand and right-hand sides of this identity, using the property of the modulus of a sum and inequality (5.58), we get the following relations*:

$$4|\operatorname{Re}(Ax, y)| \leqslant \mu\|x+y\|^2+\mu\|x-y\|^2=2\mu(\|x\|^2+\|y\|^2).$$

From this, for $\|x\|=\|y\|=1$, we get an inequality

$$|\operatorname{Re}(Ax, y)| \leqslant \mu.$$

Setting $y=Ax/\|Ax\|$ in this inequality (it is evident that $\|y\|=1$) and bearing in mind that the number $(Ax, Ax)=\|Ax\|^2$ is real (and consequently, $\operatorname{Re}(Ax, Ax)=(Ax, Ax)=\|Ax\|^2$), we obtain $\|Ax\| \leqslant \mu$, $\|x\|=1$. From this, in accordance with inequality (5.53), we find that $\|A\| \leqslant \mu$.

To complete the proof, we must compare the inequality obtained with inequality (5.57) and use the definition of the number μ (see (5.56)).

5.5.4. Other properties of self-adjoint operators. We shall prove here a number of significant properties of linear operators connected with the concept of a norm. We shall first establish the necessary and sufficient condition for an operator to be self-adjoint. Let us prove the following theorem.

Theorem 5.18. *For the linear operator A to be self-adjoint, it is necessary and sufficient that*

$$\operatorname{Im}(Ax, x)=0^{**}.$$

Proof. By Theorem 5.13, the arbitrary linear operator A can be represented as $A=A_R+iA_I$, where A_R and A_I are self-adjoint opera-

* We have used the definition of the norm of an element in a complex Euclidear space.

** The symbol $\operatorname{Im}(Ax, x)$ designates the imaginary part of the complex numbei (Ax, x). The equality $\operatorname{Im}(Ax, x)=0$ means that the number (Ax, x) is real.

ors. Therefore,

$$(\boldsymbol{Ax},\ \boldsymbol{x})=(\boldsymbol{A}_R\boldsymbol{x},\ \boldsymbol{x})+i(\boldsymbol{A}_I\boldsymbol{x},\ \boldsymbol{x}),$$

nd, in accordance with Theorem (5.15), the numbers $(\boldsymbol{A}_R\boldsymbol{x},\ \boldsymbol{x})$ and $(\boldsymbol{A}_I\boldsymbol{x},\ \boldsymbol{x})$ are real for any $\boldsymbol{x}$. Consequently, these numbers are equal, espectively, to the real and the imaginary part of the complex number $(\boldsymbol{Ax},\ \boldsymbol{x})$:

$$\operatorname{Re}(\boldsymbol{Ax},\ \boldsymbol{x})=(\boldsymbol{A}_R\boldsymbol{x},\ \boldsymbol{x}),\ \operatorname{Im}(\boldsymbol{Ax},\ \boldsymbol{x})=(\boldsymbol{A}_I\boldsymbol{x},\ \boldsymbol{x}).$$

We assume that $\boldsymbol{A}$ is a self-adjoint operator.

By Theorem 5.15, in this case the number $(\boldsymbol{Ax},\ \boldsymbol{x})$ is real and, herefore, $\operatorname{Im}(\boldsymbol{Ax},\ \boldsymbol{x})=0$. We have proved the necessity of the hyothesis.

Let us now prove its sufficiency.

Suppose $\operatorname{Im}(\boldsymbol{Ax},\ \boldsymbol{x})=(\boldsymbol{A}_I\boldsymbol{x},\ \boldsymbol{x})=0$. Hence $\|\boldsymbol{A}_I\|=0$, i.e. $\boldsymbol{A}_I=\boldsymbol{0}$. 'herefore, $\boldsymbol{A}=\boldsymbol{A}_R$, where $\boldsymbol{A}_R$ is a self-adjoint operator. We have thus roved the theorem.

In the following assertions we elucidate certain properties of the igenvalues of self-adjoint operators.

Lemma. *Any eigenvalue λ of the arbitrary linear self-adjoint oprator $\boldsymbol{A}$ in a Euclidean space is equal to the scalar product $(\boldsymbol{Ax},\ \boldsymbol{x})$, where $\boldsymbol{x}$ is some vector satisfying the condition $\|\boldsymbol{x}\|=1$:*

$$\lambda=(\boldsymbol{Ax},\ \boldsymbol{x}),\quad \|\boldsymbol{x}\|=1.$$

Proof. Since λ is an eigenvalue of the operator $\boldsymbol{A}$, there is a nonero vector $\boldsymbol{z}$ such that

$$\boldsymbol{Az}=\lambda\boldsymbol{z}. \tag{5.60}$$

Setting $\boldsymbol{x}=\boldsymbol{z}/\|\boldsymbol{z}\|$ (it is evident that $\|\boldsymbol{x}\|=1$), we rewrite (5.60) s follows: $\boldsymbol{Ax}=\lambda\boldsymbol{x}$, $\|\boldsymbol{x}\|=1$. Hence we get a relation $(\boldsymbol{Ax},\ \boldsymbol{x})=(\boldsymbol{x},\ \boldsymbol{x})=\lambda\|\boldsymbol{x}\|^2=\lambda$, i.e. (5.59) holds true. We have proved the emma.

Corollary. *Suppose $\boldsymbol{A}$ is a self-adjoint operator and λ is any eigenalue of that operator. Suppose furthermore that*

$$m=\inf_{\|x\|=1}(\boldsymbol{Ax},\ \boldsymbol{x}),\ M=\sup_{\|x\|=1}(\boldsymbol{Ax},\ \boldsymbol{x}). \tag{5.61}$$

The following inequalities hold true:

$$m\leqslant\lambda\leqslant M. \tag{5.62}$$

Remark 1. Since the scalar product $(\boldsymbol{Ax},\ \boldsymbol{x})$ is a continuous funcon of $\boldsymbol{x}$, this function is bounded on the closed set $\|x\|=1$ and atains its least upper bound M and greatest lower bound m (called a upremum and an infimum respectively).

Remark 2. In accordance with Theorem 5.16, the eigenvalues of self-adjoint operator are real. Inequalities (5.62) are, therefore,

Proof of the corollary. Since any eigenvalue λ satisties relation (5.59), every eigenvalue is, evidently, confined between the least upper bound M and the greatest lower bound m of the scalar product $(\boldsymbol{Ax}, \boldsymbol{x})$. Inequalities (5.62) are, therefore, valid.

We shall prove that the numbers m and M defined by relations (5.16) are, respectively, the least and the greatest value of the self-adjoint operator $\boldsymbol{A}$. We shall first verify the validity of the following statement.

Theorem 5.19 *Assume that $\boldsymbol{A}$ is a self-adjoint operator and, in addition, $(\boldsymbol{Ax}, \boldsymbol{x}) \geqslant 0$ for any $\boldsymbol{x}$. Then $\| \boldsymbol{A} \|$ is equal to the greatest eigenvalue of that operator.**

Proof. We have remarked (see the statement made in 5.5.3) that $\| \boldsymbol{A} \| = \sup\limits_{\| x \| = 1} | (\boldsymbol{Ax}, \boldsymbol{x}) |$. Since $(\boldsymbol{Ax}, \boldsymbol{x}) \geqslant 0$, we have $\| \boldsymbol{A} \| = \sup\limits_{\| x \| = 1} (\boldsymbol{Ax}, \boldsymbol{x})$. According to Remark 1 of this subsection,

$$(\boldsymbol{Ax}_0, \boldsymbol{x}_0) = \| \boldsymbol{A} \| = \lambda$$

for a certain $\boldsymbol{x}_0$, $\| \boldsymbol{x}_0 \| = 1$.

Using the definition of a norm and the equations we have just written, we obtain relations**

$$\| (\boldsymbol{A} - \lambda \boldsymbol{I}) \boldsymbol{x}_0 \|^2 = \| \boldsymbol{Ax}_0 \|^2 - 2\lambda (\boldsymbol{Ax}_0, \boldsymbol{x}_0) + \lambda^2 \cdot \| \boldsymbol{x}_0 \|^2$$

$$= \| \boldsymbol{A} \|^2 - 2 \| \boldsymbol{A} \| \cdot \| \boldsymbol{A} \| + \| \boldsymbol{A} \|^2 \cdot 1 = 0.$$

Thus, $(\boldsymbol{A} - \lambda \boldsymbol{I}) \boldsymbol{x}_0 = \boldsymbol{0}$, or, to put it otherwise, $\boldsymbol{Ax}_0 = \lambda \boldsymbol{x}_0$, i.e. $\lambda = \| \boldsymbol{A} \|$ is an eigenvalue of the operator $\boldsymbol{A}$. The fact that λ is the greatest eigenvalue follows from the corollary of the last lemma. We have thus proved the theorem.

Let us now prove that the numbers m and M (see (5.61)) are, respectively, the least and the greatest eigenvalue of the self-adjoint operator $\boldsymbol{A}$.

Theorem 5.20. *Let $\boldsymbol{A}$ be a self-adjoint operator, and m and M are, respectively, the greatest lower and the least upper bound of $(\boldsymbol{Ax}, \boldsymbol{x})$ on the set $\| \boldsymbol{x} \| = 1$. These numbers are the least and the greatest eigenvalue of the operator $\boldsymbol{A}$.*

Proof. It is, evidently, sufficient to prove that the numbers m and M are eigenvalues of the operator $\boldsymbol{A}$. Then, it follows immediately from inequalities (5.62) that m and M are, respectively, the least and the greatest eigenvalue.

We shall first prove that M is an eigenvalue. For that purpose, we

---

* Since there is a finite number of eigenvalues and they are real, we can indicate the greatest of them.

** We have also used the equality $\| \boldsymbol{Ax}_0 \|^2 = \| \boldsymbol{A} \|^2$ which follows from the relations $\| \boldsymbol{A} \| = (\boldsymbol{Ax}_0, \boldsymbol{x}_0) \leqslant \| \boldsymbol{Ax}_0 \|$ and $\| \boldsymbol{A} \| = \sup\limits_{\| x \| = 1} \| \boldsymbol{Ax} \|$.

shall consider a self-adjoint operator $\boldsymbol{B}=\boldsymbol{A}-m\boldsymbol{I}$. Since

$$(\boldsymbol{Bx},\ \boldsymbol{x})=(\boldsymbol{Ax},\ \boldsymbol{x})-m(\boldsymbol{x},\ \boldsymbol{x})\geqslant 0,$$

the operator $\boldsymbol{B}$ satisfies the conditions of Theorem 5.19 and, therefore, the norm $\|\boldsymbol{B}\|$ of that operator is equal to its greatest eigenvalue. On the other hand,

$$\|\boldsymbol{B}\|=\sup_{\|x\|=1}(\boldsymbol{Bx},\ \boldsymbol{x})=\sup_{\|x\|=1}(\boldsymbol{Ax},\ \boldsymbol{x})-m=M-m.$$

Thus, $(M-m)$ is the greatest eigenvalue of the operator $\boldsymbol{B}$. Consequently, there is nonzero vector $\boldsymbol{x}_0$ such that

$$\boldsymbol{Bx}_0=(M-m)\boldsymbol{x}_0. \tag{5.63}$$

Since $\boldsymbol{B}=\boldsymbol{A}-m\boldsymbol{I}$, we have $\boldsymbol{Bx}_0=\boldsymbol{Ax}_0-m\boldsymbol{Ix}_0=\boldsymbol{Ax}_0-m\boldsymbol{x}_0$. Substituting the expression $\boldsymbol{Bx}_0$ into left-hand side of (5.63), we get, after simple transformations, a relation $\boldsymbol{Ax}_0=M\boldsymbol{x}_0$. Thus, M is an eigenvalue of the operator $\boldsymbol{A}$.

Let us now make sure that the number m is also an eigenvalue of the operator $\boldsymbol{A}$.

We shall consider the self-adjoint operator $\boldsymbol{B}=-\boldsymbol{A}$. Evidently, $-m=\sup\limits_{\|x\|=1}(\boldsymbol{Bx},\ \boldsymbol{x})$. According to the proof we have just carried out, the number $-m$ is an eigenvalue of the operator $\boldsymbol{B}$. Since $\boldsymbol{B}=-\boldsymbol{A}$, m is an eigenvalue of the operator $\boldsymbol{A}$. We have proved the theorem.

In the following theorem we elucidate a significant property of the eigenvectors of a self-adjoint operator.

Theorem 5.21. *Every self-adjoint linear operator $\boldsymbol{A}$ acting in an n-dimensional Euclidean space V possesses n linearly independent pairwise orthogonal unit eigenvectors.*

Proof. Let λ_1 be the maximum eigenvalue of the operator $\boldsymbol{A}$ ($\lambda_1=\sup\limits_{\|x\|=1}(\boldsymbol{Ax},\ \boldsymbol{x})$). We denote by $\boldsymbol{e}_1$ the eigenvector corresponding to λ_1 and satisfying the condition $\|\boldsymbol{e}_1\|=1$ (the possibility of choosing it follows from the proof of the lemma in this subsection).

We denote by V_1 an $(n-1)$-dimensional subspace of the space V which is orthogonal to $\boldsymbol{e}_1$. Evidently, V_1 is an invariant subspace of the operator $\boldsymbol{A}$ (i.e. if $\boldsymbol{x}\in V_1$ then $\boldsymbol{Ax}\in V_1$ as well). Indeed, suppose $\boldsymbol{x}\in V_1$ (i.e. $(\boldsymbol{x},\ \boldsymbol{e}_1)=0$). Then*

$$(\boldsymbol{Ax},\ \boldsymbol{e}_1)=(\boldsymbol{x},\ \boldsymbol{Ae}_1)=\lambda_1(\boldsymbol{x},\ \boldsymbol{e}_1)=0.$$

Consequently, $\boldsymbol{Ax}$ is an element of V_1 and, therefore, V_1 is an invar-

* We have used the property of self-adjointness of the operator $(\boldsymbol{Ax},\ \boldsymbol{e}_1)=(\boldsymbol{x},\ \boldsymbol{Ae}_1)$ and the fact that $\boldsymbol{e}_1$ is an eigenvector of the operator: $\boldsymbol{Ae}_1=\lambda_1\boldsymbol{e}_1$.

iant subspace of the operator $\boldsymbol{A}$. This enables us to consider the operator $\boldsymbol{A}$ in the subspace V_1. In this subspace $\boldsymbol{A}$ is a self-adjoint operator. Consequently, there is a maximum eigenvalue λ_2 of that operator which can be found by the relation*

$$\lambda_2 = \max_{\substack{\|x\|=1 \\ x \perp e_1}} (\boldsymbol{A}x, x).$$

In addition, there is a vector e_2, $e_2 \perp e_1$, $\| e_2 \| = 1$ such that $\boldsymbol{A}e_2 = \lambda_2 e_2$.

Considering then an $(n-2)$-dimensional subspace V_2 which is orthogonal to the vectors e_1 and e_2 and repeating the arguments presented above, we construct an eigenvector e_3, $\| e_3 \| = 1$ which is orthogonal to e_1 and e_2. Continuing with the same arguments, we successively find n mutually orthogonal eigenvectors $e_1, e_2, \ldots, e_n$ satisfying the condition $\| e_i \| = 1$, $i = 1, 2, \ldots, n$.

Remark 1. In what follows we agree to enumerate the eigenvalues of a self-adjoint operator in a decreasing order, with the repeating, i.e. multiple, eigenvalues taken into account. In this case, $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n$ and the eigenvectors $e_1, e_2, \ldots, e_n$ corresponding to them can be considered mutually orthogonal and satisfying the condition $\| e_i \| = 1$. Thus,

$$(e_i, e_j) = \begin{cases} 1 & \text{for } i = j, \\ 0 & \text{for } i \neq j. \end{cases}$$

Remark 2. It follows from the arguments presented in the proof of Theorem 5.21 that $\lambda_{m+1} = \max\limits_{\substack{x \perp e_k \\ k=1,2,\ldots,m}} \dfrac{(\boldsymbol{A}x, x)}{(x, x)}$. This relation can also be written in the form $\lambda_{m+1} = \max\limits_{x \perp E_m} \dfrac{(\boldsymbol{A}x, x)}{(x, x)}$, where E_m is a linear closure of the vectors $e_1, e_2, \ldots, e_m$. The validity of the remark follows from the fact that $(x, x) = \| x \|^2$ and, therefore

$$\frac{(\boldsymbol{A}x, x)}{(x, x)} = \left(A \frac{x}{\| x \|}, \frac{x}{\| x \|} \right),$$

with the norm of the element $x/\| x \|$ being equal to unity.

Let $\mathcal{E}_m$ be the set of all m-dimensional subspaces of the space V. The following significant **minimax** property of eigenvalues holds true.

Theorem 5.22. *Suppose $\boldsymbol{A}$ is a self-adjoint operator and* $\lambda_1, \lambda_2, \ldots$

* The symbol $e_1 \perp e_2$ designates the orthogonality of the vectors e_1 and e_2.

..., λ_n are its eigenvalues enumerated in the order indicated in Remark 1. Then

$$\lambda_{m+1} = \min_{E \in \mathcal{E}_m} \max_{x \perp E} \frac{(Ax, x)}{(x, x)}. \tag{5.64}$$

Proof. Suppose E_m is a linear closure of the eigenvectors $e_1, e_2, \ldots, e_m$ of the operator A (see Remark 1). By virtue of Remark 2

$$\max_{x \perp E_m} \frac{(Ax, x)}{(x, x)} = \lambda_{m+1}.$$

Therefore, to prove the theorem, it is sufficient to verify the validity of the relation

$$\max_{x \perp E \subset \mathcal{E}_m} \frac{(Ax, x)}{(x, x)} \geqslant \max_{x \in E_m} \frac{(Ax, x)}{(x, x)} = \lambda_{m+1} \tag{5.65}$$

for any $E \in \mathcal{E}_m$.

Let us prove relation (5.65).

We denote by $E^{\perp}$ the orthogonal complement of the space E (see 4.2.3). It follows from Theorem 2.10 that the dimension of $E^{\perp}$ is $n-m$. Consequently,

$$\dim E^{\perp} + \dim E_{m+1} = (n-m) + (m+1) = n+1 > n.$$

By Theorem 2.9 this means that the intersection of the subspaces $E^{\perp}$ and E_{m+1} contains a nonzero element. Thus, there exists an element $\tilde{x}$ such that $\tilde{x} \perp E$, $\| \tilde{x} \| = 1$, $\tilde{x} \in E_{m+1}$, i.e. $x = \sum_{k=1}^{m+1} c_k e_k$. Since $\| \tilde{x} \| = 1$ and the basis $e_1, e_2, \ldots, e_{m+1}$ is orthonormal it follows, by virtue of the Pythagoras Theorem (see 4.1.2), that

$$\| \tilde{x} \|^2 = \sum_{k=1}^{m+1} | c_k |^2 = 1. \tag{5.66}$$

We also have $A\tilde{x} = A \sum_{k=1}^{m+1} c_k e_k = \sum_{k=1}^{m+1} c_k A e_k$. Since e_k are eigenvectors of the operator A, the last relations yield $A\tilde{x} = \sum_{k=1}^{m+1} c_k \lambda_k e_k$.

From this and from the orthonormality of e_k we find that the relation

$$(A\tilde{x}, x) = \left(\sum_{k=1}^{m+1} c_k \lambda_k e_k, \sum_{p=1}^{m+1} c_p e_p \right) = \sum_{k=1}^{m+1} | c_k |^2 \lambda_k \tag{5.67}$$

holds true.

We have enumerated the eigenvalues in a decreasing order with their possible multiplicity taken into account. Therefore, $\lambda_{m+1} \geqslant \lambda_k$, $k=1, 2, \ldots, m$. From this and from (5.67) and (5.66) we get

$$(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}) = \sum_{k=1}^{m+1} |c_k|^2 \lambda_k \geqslant \lambda_{m+1} \sum_{k=1}^{m+1} |c_k|^2 = \lambda_{m+1}.$$

Noting that the norm of the element $\boldsymbol{x}/\|\boldsymbol{x}\|$ is equal to 1 and $\|\tilde{\boldsymbol{x}}\| = 1$ for any $\boldsymbol{x} \neq 0$ and taking into account that $\tilde{\boldsymbol{x}} \perp E$, we get

$$\max_{\boldsymbol{x} \perp E} \frac{(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{x})}{(\boldsymbol{x}, \boldsymbol{x})} = \max_{\boldsymbol{x} \perp E} \left(\boldsymbol{A} \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}, \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} \right) \geqslant (\boldsymbol{A}\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}}) \geqslant \lambda_{m+1}.$$

We have thus established relations (5.65) and proved the theorem.

5.5.5. Expansion of self-adjoint operators in terms of eigenfunctions. Hamilton-Cayley theorem. Let us consider a self-adjoint operator $\boldsymbol{A}$ and its eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n$. In this case $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ is an orthonormal basis consisting of eigenvectors corresponding to $\{\lambda_i\}$. Let $\boldsymbol{x} \in V$. Then

$$\boldsymbol{x} = \sum_{k=1}^{n} (\boldsymbol{x}, \boldsymbol{e}_k) \boldsymbol{e}_k \tag{5.68}$$

(see 4.2.3), and since $\boldsymbol{A}\boldsymbol{e}_k = \lambda_k \boldsymbol{e}_k$, we get, by means of (5.68).

$$\boldsymbol{A}\boldsymbol{x} = \sum_{k=1}^{n} \lambda_k (\boldsymbol{x}, \boldsymbol{e}_k) \boldsymbol{e}_k. \tag{5.69}$$

The operator $\boldsymbol{P}_k$, defined by the relation

$$\boldsymbol{P}_k \boldsymbol{x} = (\boldsymbol{x}, \boldsymbol{e}_k) \boldsymbol{e}_k, \tag{5.70}$$

is known as a **projection** onto the one-dimensional subspace generated by the vectors $\boldsymbol{e}_k$.

It immediately follows from the properties of a scalar product that $\boldsymbol{P}_k$ is a self-adjoint lineaar operator.

Note the following significant properties of projections:

1°. $\boldsymbol{P}_k^2 = \boldsymbol{P}_k$ (hence $\boldsymbol{P}_k^m = \boldsymbol{P}_k$, where m is natural).

2°. $\boldsymbol{P}_k \boldsymbol{P}_j = 0$, where $k \neq j$.

The proof of these properties follows from the relations

$$(\boldsymbol{P}_k \boldsymbol{P}_j) \boldsymbol{x} = \boldsymbol{P}_k (\boldsymbol{P}_j \boldsymbol{x}) = \boldsymbol{P}_k (\boldsymbol{x}, \boldsymbol{e}_j) \boldsymbol{e}_j$$

$$= (\boldsymbol{x}, \boldsymbol{e}_j)(\boldsymbol{e}_j, \boldsymbol{e}_k) \boldsymbol{e}_k = \begin{cases} (\boldsymbol{x}, \boldsymbol{e}_k) \boldsymbol{e}_k & \text{for } k = j, \\ O & \text{for } k \neq j. \end{cases}$$

Note that it follows immediately from definition (5.70) that $\boldsymbol{P}_k$ commutes with every operator which commutes with $\boldsymbol{A}$.

From relations (5.68), (5.69) and (5.70) we get the following expressions for $\boldsymbol{x}$ and $\boldsymbol{Ax}$:

$$\boldsymbol{x}=\sum_{k=1}^{n}\boldsymbol{P}_k\boldsymbol{x}, \tag{5.71}$$

$$\boldsymbol{Ax}=\sum_{k=1}^{n}\lambda_k\boldsymbol{P}_k\boldsymbol{x}. \tag{5.72}$$

It follows from equation (5.71) that the operator $\sum_{k=1}^{n}\boldsymbol{P}_k$ is *identical:*

$$I=\sum_{k=1}^{n}\boldsymbol{P}_k. \tag{5.73}$$

Equation (5.72) yields a so-called **expansion of a self-adjoint operator in terms of eigenfunctions:**

$$\boldsymbol{A}=\sum_{k=1}^{n}\lambda_k\boldsymbol{P}_k. \tag{5.74}$$

Properties 1° and 2° of projections and relation (5.74) yield the following expression for $\boldsymbol{A}^2$:

$$\boldsymbol{A}^2=\sum_{k=1}^{n}\lambda_k^2\boldsymbol{P}_k.$$

It is evident that in general

$$\boldsymbol{A}^s=\sum_{k=1}^{n}\lambda_k^s\boldsymbol{P}_k \tag{5.75}$$

for any integral positive s.

Let us consider an arbitrary polynomial $p(\lambda)=\sum_{i=1}^{n}c_i\lambda^i$. By definition, it is assumed that $p(\boldsymbol{A})=\sum_{k=1}^{m}c_k\boldsymbol{A}^k$. Paying attention to relation (5.75), it is easy to get the foollowing expression for $p(\boldsymbol{A})$:

$$p(\boldsymbol{A})=\sum_{i=1}^{m}p(\lambda_i)\boldsymbol{P}_i. \tag{5.76}$$

Let us prove the following theorem.

Theorem 5.23. (Hamilton-Cayley Theorem). *If $\boldsymbol{A}$ is a self-adjoint operator and $p(\lambda)=\det(\boldsymbol{A}-\lambda\boldsymbol{I})$ is a characteristic polynomial of that operator, then*

$$p(\boldsymbol{A})=0.$$

Proof. Indeed, if $\boldsymbol{A}$ is a self-adjoint operator and λ_i are its eigen-

values, then, in accordance with Theorem 5.8, λ_i is a root of the characteristic equation, i.e. $p(\lambda_i)=0$. From this and from (5.76) it follows that $p(\boldsymbol{A})=0$. We have proved the theorem.

5.5.6. Positive operators. The *m*th-degree roots of an operator. The self-adjoint operator $\boldsymbol{A}$ is said to be **positive** if the relation

$$(\boldsymbol{Ax},\ \boldsymbol{x})\geqslant 0 \tag{5.77}$$

holds for any $\boldsymbol{x}$ from V.

If the operator $\boldsymbol{A}$ is positive and it follows from the condition $(\boldsymbol{Ax},\ \boldsymbol{x})=0$ that $\boldsymbol{x}=0$, then $\boldsymbol{A}$ is said to be a **positive definite operator.**

Positive and positive definite operators are designated as $\boldsymbol{A}\geqslant 0$ and $\boldsymbol{A}>0$ respectively.

Note the following simple *statement.*

Every eigenvalue of a positive (positive definite) operator is non-negative (positive).

This statement follows from simple arguments.

Let λ be an eigenvalue of the operator $\boldsymbol{A}$. Then, in accordance with the lemma in 5.5.4, there is an element $\boldsymbol{x}$, $\|\boldsymbol{x}\|=1$, such that $\lambda=(\boldsymbol{Ax},\ \boldsymbol{x})$.

From this and from (5.77) we find that $\lambda\geqslant 0$ for positive operators and $\lambda>0$ for positive definite operators. We have proved the statement.

We introduce the concept of an ***m*th-degree root** (m being a natural number) **of an operator.**

Definition. *A root of degree m of the operator* $\boldsymbol{A}$ *is an operator* $\boldsymbol{B}$ *such that* $\boldsymbol{B}^m=\boldsymbol{A}$.

A root of degree m of the operator $\boldsymbol{A}$ is symbolized as $\boldsymbol{A}^{1/m}$.

It is natural to isolate a class of operators for which the operation of finding an mth-degree root would have sense. The following theorem provides a definite answer to this question.

Theorem 5.24. *Let* $\boldsymbol{A}$ *be a positive self-adjoint operator* $\boldsymbol{A}\geqslant 0$. *Then there is a positive self-adjoint operator* $\boldsymbol{A}^{1/m}$, $\tilde{\boldsymbol{A}}^{1/m}\geqslant 0$, *for any natural m.*

Proof. We denote by λ_k the eigenvalues of the operator $\boldsymbol{A}$, and let $\{\boldsymbol{e}_k\}$ be an orthonormal basis of the eigenvectors. Next we denote by $\boldsymbol{P}_k$ the projection onto a one-dimensional subspace generated by $\boldsymbol{e}_k$.

In accordance with what was said in the preceding subsection, we have an expansion (5.74) of the self-adjoint operator $\boldsymbol{A}$ in terms of eigenfunctions:

$$\boldsymbol{A}=\sum_{k=1}^{n}\lambda_k\boldsymbol{P}_k. \tag{5.74}$$

Since $\lambda_k\geqslant 0$ (see the statement we have just proved), we can introduce the following self-adjoint operator $\boldsymbol{B}$:

$$B=\sum_{k=1}^{n}\lambda_k^{1/m}P_k. \tag{5.78}$$

In accordaance with (5.70), there holds a relation

$$(P_k x,\ x)\geqslant 0$$

which yields the positivity of the operators P_k and that of the operator B (see (5.78)).

It follows from properties 1° and 2° of the projections P_k (see 5.5.5) that $B^m=\sum_{k=1}^{n}\lambda_k P_k$. Comparing this expression for B^m with expression (5.74) for A, we find that $B^m=A$. We have found above that the operator B is positive. And this is what we wished to prove.

Remark 1. We note without proof that there exists a unique positive operator $A^{1/m}$.

Remark 2. In the orthonormal basis $\{e_k\}$ of the eigenvectors of the operator A the matrix of the operator $A^{1/m}$ has the following form:

$$\begin{pmatrix} \lambda_1^{1/m} & 0 & & 0 \\ 0 & \lambda_2^{1/m} & & 0 \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \dots & \lambda_n^{1/m} \end{pmatrix}$$

5.6. Reducing a Quadratic Form to the Sum of Squares

We shall discuss here the problem of choosing a basis in which a quadratic form (an invariant quadratic function of the coordinates of a vector; the exact definition will be given later on) has the simplest form.

Quadratic forms will be studied in Chapter 7. In particular, we shall consider there various methods of reducing quadratic forms to a sum of squares.

We introduce the concept of the so-called **Hermitian forms.**

Definition. *The sesquilinear form* $B(x,\ y)$ is called ***Hermitian*** *if for any* x *and* y *there holds a relation*

$$B(x,\ y)=\overline{B(x,\ y)}. \tag{5.79}$$

In accordance with the corollary of Theorem 5.11, any sesquilinear form $B(x,\ y)$ (a Hermitian form inclusive) can be uniquely represented as

$$B(x,\ y)=(Ax,\ y), \tag{5.80}$$

where A is a linear operator.

Let us prove the following two statements in which we shall elucidate the conditions under which a sesquilinear form is Hermitian.

Theorem 5.25. *For the sesquilinear form $B(\boldsymbol{x}, \boldsymbol{y})$ to be Hermitian, it is necessary and sufficient that the operator $\boldsymbol{A}$ in representation* (5.80) *of this form should be self-adjoint* ($\boldsymbol{A}=\boldsymbol{A}^*$).

Proof. Indeed, if $\boldsymbol{A}$ is a self-adjoint operator, then, using the properties of a scalar product, we obtain

$$B(\boldsymbol{x}, \boldsymbol{y})=(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{y})=(\boldsymbol{x}, \boldsymbol{A}\boldsymbol{y})=\overline{(\boldsymbol{A}\boldsymbol{y}, \boldsymbol{x})}=\overline{B(\boldsymbol{y}, \boldsymbol{x})}.$$

Thus condition (5.79) is fulfilled, i.e. the form $B(\boldsymbol{x}, \boldsymbol{y})=(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{y})$ is Hermitian.

And if the form $B(\boldsymbol{x}, \boldsymbol{y})=(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{y})$ is Hermitian, we can again take the properties of a scalar product into account and get

$$(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{y})=B(\boldsymbol{x}, \boldsymbol{y})=\overline{B(\boldsymbol{y}, \boldsymbol{x})}=\overline{(\boldsymbol{A}\boldsymbol{y}, \boldsymbol{x})}=(\boldsymbol{x}, \boldsymbol{A}\boldsymbol{y}).$$

Thus we have $(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{y})=(\boldsymbol{x}, \boldsymbol{A}\boldsymbol{y})$, i.e. the operator $\boldsymbol{A}$ is self-adjoint. And this is what we set out to prove.

Theorem 5.26. *For the sesquilinear form $B(\boldsymbol{x}, \boldsymbol{y})$ to be Hermitian, it is necessary and sufficient that the function $B(\boldsymbol{x}, \boldsymbol{x})$ should be real.*

Proof. The form $B(\boldsymbol{x}, \boldsymbol{y})$ is Hermitian if and only if the linear operator $\boldsymbol{A}$ in representation (5.80) of this form is self-adjoint (see Theorem 5.25). But in accordance with Theorem 5.18, for the operator $\boldsymbol{A}$ to be self-adjoint, it is necessary and sufficient that the scalar product $(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{x})$ should be real for any $\boldsymbol{x}$. We have proved the theorem.

We shall now introduce the concept of a **quadratic form.**

Let $B(\boldsymbol{x}, \boldsymbol{y})$ be a Hermitian form.

*The **quadratic form** corresponding to the form $B(\boldsymbol{x}, \boldsymbol{y})$ is a function $B(\boldsymbol{x}, \boldsymbol{x})$.*

We shall now prove the following *theorem on the reduction of a quadratic form to a sum of squares* (*to a diagonal form*).

Theorem 5.27. *Suppose $B(\boldsymbol{x}, \boldsymbol{y})$ is a Hermitian form defined on various vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ of the n-dimensional Euclidean space V. Then this space contains an orthonormal basis $\{\boldsymbol{e}_k\}$ and there are real numbers λ_k such that for any $\boldsymbol{x}$ belonging to V the quadratic form $B(\boldsymbol{x}, \boldsymbol{x})$ can be represented as the following sum of the squares of the coordinates ξ_k of the vector $\boldsymbol{x}$ in the basis $\{\boldsymbol{e}_k\}$:*

$$B(\boldsymbol{x}, \boldsymbol{x})=\sum_{k=1}^{n} \lambda_k |\xi_k|^2. \tag{5.81}$$

Proof. The form $B(\boldsymbol{x}, \boldsymbol{y})$ being Hermitian, there is self-adjoint operator $\boldsymbol{A}$, in accordance with Theorem 5.25, such that

$$B(\boldsymbol{x}, \boldsymbol{y})=(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}). \tag{5.82}$$

Let us direct our attention to Theorem 5.21. By this theorem, we can indicate for the operator $\boldsymbol{A}$ an orthonormal basis $\{\boldsymbol{e}_k\}$ consisting of the eigenvectors of that operator. If λ_k are the eigenvalues of $\boldsymbol{A}$ and ξ_k are the coordinates of the vector $\boldsymbol{x}$ in the basis $\{\boldsymbol{e}_k\}$ so that

$$\boldsymbol{x}=\sum_{k=1}^{n} \xi_k \boldsymbol{e}_k, \tag{5.83}$$

then, using formula (5.12) and the relation $\boldsymbol{A}\boldsymbol{e}_k=\lambda_k \boldsymbol{e}_k$*, we get the following expression for $\boldsymbol{A}\boldsymbol{x}$:

$$\boldsymbol{A}\boldsymbol{x}=\sum_{k=1}^{n} \lambda_k \xi_k \boldsymbol{e}_k. \tag{5.84}$$

From (5.83), (5.84) and the orthonormality of the basis $\{\boldsymbol{e}_k\}$ we obtain the following expression for $(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{x})$:

$$(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{x})=\sum_{k=1}^{n} \lambda_k |\xi_k|^2.$$

From this expression and from relation (5.82) we get (5.81). We have thus proved the theorem.

Let us now prove an important theorem on simultaneous reduction of two quadratic forms to a sum of squares.

Theorem 5.28. *Suppose $A(\boldsymbol{x}, \boldsymbol{y})$ and $B(\boldsymbol{x}, \boldsymbol{y})$ are Hermitian forms defined on various vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ of the n-dimensional linear space V. We assume that an inequality $B(\boldsymbol{x}, \boldsymbol{x})>0$ holds for all nonzero elements $\boldsymbol{x}$ belonging to V. Then there is a basis $\{\boldsymbol{e}_k\}$ in the space V such that the quadratic forms $A(\boldsymbol{x}, \boldsymbol{x})$ and $B(\boldsymbol{x}, \boldsymbol{x})$ can be represented as*

$$A(\boldsymbol{x}, \boldsymbol{x})=\sum_{k=1}^{n} \lambda_k |\xi_k|^2, \tag{5.85}$$

$$B(\boldsymbol{x}, \boldsymbol{x})=\sum_{k=1}^{n} |\xi_k|^2, \tag{5.86}$$

where λ_k are real numbers and ξ_k are the coordinates of the vector $\boldsymbol{x}$ in the basis $\{\boldsymbol{e}_k\}$.

Proof. Since under an additional requirement that $B(\boldsymbol{x}, \boldsymbol{x})>0$ for $\boldsymbol{x}\neq 0$ the properties of a scalar product and those of the Hermitian form $B(\boldsymbol{x}, \boldsymbol{y})$ are similarly formulated, we can introduce a scalar product $(\boldsymbol{x}, \boldsymbol{y})$ of vectors in the linear space V setting

$$(\boldsymbol{x}, \boldsymbol{y})=B(\boldsymbol{x}, \boldsymbol{y}). \tag{5.87}$$

* This relation follows from the fact that λ_k and $\boldsymbol{e}_k$ are the eigenvalues and eigenvectors of the operator $\boldsymbol{A}$ respectively.

Thus V is a Euclidean space with the scalar product (5.87).

By Theorem 5.27, we can indicate in V an orthonormal basis $\{e_k\}$ and real numbers λ_k such that in this basis the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ is represented as (5.85).

On the other hand, in any orthonormal basis the scalar product $(\boldsymbol{x}, \boldsymbol{x})$, equal to $B(\boldsymbol{x}, \boldsymbol{x})$ in accordance with (5.87), is equal to the sum of squares of the absolute values of the coordinates of the vector $\boldsymbol{x}$. Thus, the representation of $B(\boldsymbol{x}, \boldsymbol{x})$ in the form (5.86) is also well-founded. And this is what we had to prove.

5.7. Unitary and Normal Operators

We discuss here the properties of a significant class of operators acting in the Euclidean space V.

Definition 1. *The linear operator $\boldsymbol{U}$ from $L(V, V)$ is called* ***unitary*** *if for any elements $\boldsymbol{x}$ and $\boldsymbol{y}$ belonging to V there holds a relation*

$$(\boldsymbol{Ux}, \boldsymbol{Uy})=(\boldsymbol{x}, \boldsymbol{y}). \tag{5.88}$$

In what follows, relation (5.88) is called the **condition under which an operator is unitary.**

Remark 1. It follows from condition (5.88) under which an operator is unitary that for any unitary operator $\boldsymbol{U}$ the equality $\| \boldsymbol{Ux} \| = \| \boldsymbol{x} \|$ holds true.

Note the following *assertion.*

If λ is an eigenvalue of the unitary operator $\boldsymbol{U}$, then $|\lambda|=1$.

Indeed, if λ is an eigenvalue of $\boldsymbol{U}$, then there is an element $\boldsymbol{e}$ such that $\| \boldsymbol{e} \|=1$ and $\boldsymbol{Ue}=\lambda\boldsymbol{e}$. From this and from Remark 1 it follows that $|\lambda| = \| \lambda\boldsymbol{e} \| = \| \boldsymbol{Ue} \| = \| \boldsymbol{e} \| = 1$. We have proved the assertion.

Let us now prove the following theorem.

Theorem 5.29. *For the linear operator $\boldsymbol{U}$, acting in the Euclidean space V, to be unitary, it is necessary and sufficient that the relation*

$$\boldsymbol{U}^*=\boldsymbol{U}^{-1} \tag{5.89}$$

hold true.

Proof. **(1) Necessity.** Let the operator $\boldsymbol{U}$ be unitary, i.e. condition (5.88) be fulfilled. Taking into account the definition of an adjoint operator $\boldsymbol{U}^*$, we can rewrite this condition in the following form*.

$$(\boldsymbol{U}^*\boldsymbol{Ux}, \boldsymbol{y})=(\boldsymbol{x}, \boldsymbol{y}), \tag{5.90}$$

or, to put it otherwise, the equality

$$((\boldsymbol{U}^*\boldsymbol{U}-\boldsymbol{I})\boldsymbol{x}, \boldsymbol{y})=0$$

is satisfied for any $\boldsymbol{x}$ and $\boldsymbol{y}$.

---

* Recall that the operator $\boldsymbol{U}^*$ is said to be an adjoint of the operator $\boldsymbol{U}$ if for any $\boldsymbol{z}$ and $\boldsymbol{y}$ the relation $(\boldsymbol{z}, \boldsymbol{Uy})=(\boldsymbol{U}^*\boldsymbol{z}, \boldsymbol{y})$ is satisfied. Setting $\boldsymbol{z}=\boldsymbol{Ux}$, we get (5.90).

Fixing any element x in this equation and considering y to be arbitrary, we find that the linear operator U^*U-I acts according to the rule $(U^*U-I)x=0$.

Consequently, $U^*U=I$. We can make sure by analogy that $UU^*=I$.

Thus U and U^* are operators inverse of each other, i.e. relation (5.89) is satisfied. We have proved the necessity of the theorem.

(2) Sufficiency. Suppose condition (5.89) is fulfilled. Then, evidently, $UU^*=U^*U=I$. Directing our attention to the definition of an adjoint operator and using the relation we have just written, we get, for any x and y, relations

$$(Ux,\ Uy)=(x,\ U^*Uy)=(x,\ Iy)=(x,\ y).$$

Thus condition (5.88) under which the operator is unitary is fulfilled. Consequently, the operator U is unitary. We have thus proved the theorem.

Remark 2. While proving the theorem, we have found that condition (5.88) under which the operator U is unitary and the condition

$$U^*U=UU^*=I \tag{5.91}$$

are equivalent. We can thus base the definition of a unitary operator on condition (5.91).

This condition can also be called the **condition under which the operator U is unitary.**

We shall now introduce the concept of a **normal** operator.

Definition 2. *The linear operator A is called **normal** if the relation*

$$A^*A=AA^* \tag{5.92}$$

holds true.

Considering condition (5.91) under which an operator is unitary and condition (5.92), we see that *any unitary operator is a normal operator.*

We shall need the following auxiliary assertion.

Lemma. *Let A be a normal operator. Then the operator A and the operator A^* have a common element e such that $\|e\|=1$ and the relations $Ae=\lambda e$* and *$A^*e=\lambda e$ hold true.*

Proof. Suppose λ is an eigenvalue of the operator A and $R_\lambda=\ker(A-\lambda I)$. In other words, R_λ is a set of all elements x such that $Ax-\lambda x=0$.

Let us verify that if x belongs to R_λ then A^*x also belongs to R_λ. Indeed, if $Ax=\lambda x$ (i.e. $x\in R_\lambda$), then

$$A(A^*x)=A^*(Ax)=A^*(\lambda x)=\lambda(A^*x)$$

since A is a normal operator.

To put it otherwise, the vector A^*x is an eigenvector of the operator A and corresponds to the eigenvalue λ, i.e. belongs to R_λ.

Considering the operator A^* as an operator from R_λ into R_λ and using the inference from the corollary of Theorem 5.8 stating that every linear operator has an eigenvalue, we can assert that there is an element e in R_λ such that $\| e \| = 1$ and the relations $A^*e = \mu e$ and $Ae = \lambda e$ hold true.

Using these relations and the condition $\| e \| = 1$, we find that $(Ae, e) = (\lambda e, e) = \lambda \| e \|^2 = \lambda$, $(e, A^*e) = (e, \mu e) = \bar{\mu} \| e \|^2 = \bar{\mu}$. Since $(Ae, e) = (e, A^*e)$, it evidently follows that $\lambda = \bar{\mu}$. We have proved the lemma.

Let us now prove the following theorem.

Theorem 5.30. *Let A be a normal operator. Then there is an orthonormal basis $\{e_k\}$ consisting of the eigenelements of the operators A and A^*.*

Proof. In accordance with the lemma we have just proved, the operators A and A^* have a common eigenelement e_1, belonging to V, with $\| e_1 \| = 1$. The eigenvalues of the operators A and A^* corresponding to e_1 are equal to λ_1 and $\bar{\lambda}_1$ respectively.

Suppose V_1 is an orthogonal complement of the element e_1 to complete the space V. In other words, V_1 is the collection of all x satisfying the condition $(x, e_1) = 0$.

We shall prove that if x belongs to V_1, then Ax and A^*x belong to V_1. Indeed, if $(x, e_1) = 0$, then

$$(Ax, e_1) = (x, A^*e_1) = (x, \bar{\lambda}_1 e_1) = \lambda_1 (x, e_1) = 0,$$

i.e. $Ax \in V_1$. By analogy, if $(x, e_1) = 0$, then

$$(A^*x, e_1) = (x, Ae_1) = (x, \lambda_1 e_1) = \bar{\lambda}_1 (x, e_1) = 0,$$

i.e. $A^*x \in V_1$.

Thus V_1 is an invariant subspace of the operators A and A^*. Therefore, by the lemma we have just proved, there is an eigenelement e_2 of the operators A and A^* in the subspace V_1 such that $\| e_2 \| = 1$, $Ae_2 = \lambda_2 e_2$, $A^*e_2 = \bar{\lambda}_2 e_2$.

Next we designate as V_2 an orthogonal complement of the element e_2 to complete V_1. Reasoning as above, we shall prove that there is an eigenelement e_3 of the operators A and A^* in V_2 such that $\| e_3 \| = 1$. Continuing with similar arguments, we shall, evidently, construct, in the space V, an orthonormal basis $\{e_k\}$ consisting of the eigenelements of the operators A and A^*. We have proved the theorem.

Corollary 1. *Let A be a normal operator. There is a basis $\{e_k\}$ in which A has a diagonal matrix.*

Indeed, in accordance with the theorem we have proved, there is a basis $\{e_k\}$ consisting of the eigenvectors of the operator A. By Theorem 5.9, the matrix of the operator A is diagonal in this basis.

Corollary 2. *A unitary operator has a complete orthonormal system of eigenvectors.*

Here is the reciprocal of Theorem 5.30.

Theorem 5.31. *If the operator $\boldsymbol{A}$ acting in the n-dimensional Euclidean space V has n pairwise orthogonal eigenelements $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$, then the operator $\boldsymbol{A}$ is normal.*

Proof. Suppose $\{\boldsymbol{e}_k\}$ are pairwise orthogonal eigenvectors of the operator $\boldsymbol{A}$. Then $\boldsymbol{A}\boldsymbol{e}_k=\lambda_k\boldsymbol{e}_k$ and, according to Theorem (5.69), there holds the following representation of the operator $\boldsymbol{A}$*:

$$\boldsymbol{A}\boldsymbol{x}=\sum_{k=1}^{n}\lambda_k(\boldsymbol{x},\ \boldsymbol{e}_k)\,\boldsymbol{e}_k.$$

We shall prove that the adjoint operator $\boldsymbol{A}^*$ acts according to the rule

$$\boldsymbol{A}^*\boldsymbol{y}=\sum_{k=1}^{n}\bar{\lambda}_k(\boldsymbol{y},\ \boldsymbol{e}_k)\,\boldsymbol{e}_k. \tag{5.93}$$

It is sufficient to prove that the equality

$$(\boldsymbol{x},\ \boldsymbol{A}^*\boldsymbol{y})=(\boldsymbol{A}\boldsymbol{x},\ \boldsymbol{y}) \tag{5.94}$$

is valid for the operators $\boldsymbol{A}$ and $\boldsymbol{A}^*$ defined by relations (5.69) and (5.93).

Substituting the expression $\boldsymbol{A}^*\boldsymbol{y}$ into the left-hand side of this equation in accordance with formula (5.93), we get, after simple transformations,

$$(\boldsymbol{x},\ \boldsymbol{A}^*\boldsymbol{y})=\sum_{k=1}^{n}(\boldsymbol{x},\ \bar{\lambda}_k(\boldsymbol{y},\ \boldsymbol{e}_k)\,\boldsymbol{e}_a)=\sum_{k=1}^{n}\lambda_k\overline{(\boldsymbol{y},\ \boldsymbol{e}_k)}(\boldsymbol{x},\ \boldsymbol{e}_k)$$
$$=\sum_{k=1}^{n}\lambda_k(\boldsymbol{x},\ \boldsymbol{e}_k)(\boldsymbol{e}_k,\ \boldsymbol{y})=(\boldsymbol{A}\boldsymbol{x},\ \boldsymbol{y}).$$

We have thus proved equality (5. 94) and, therefore, the operator $\boldsymbol{A}^*$ acting according to rule (5.93) is an adjoint of the operator $\boldsymbol{A}$.

To complete the proof of the theorem, we must verify the validity of equality (5.92): $\boldsymbol{A}^*\boldsymbol{A}=\boldsymbol{A}\boldsymbol{A}^*$.

In accordance with (5.93)**, we have

$$\boldsymbol{A}\boldsymbol{A}^*\boldsymbol{x}=\sum_{k=1}^{n}\bar{\lambda}_k(\boldsymbol{x},\ \boldsymbol{e}_k)\,\boldsymbol{A}\boldsymbol{e}_k$$
$$=\sum_{k=1}^{n}\lambda_k\bar{\lambda}_k(\boldsymbol{x},\ \boldsymbol{e}_k)\,\boldsymbol{e}_k=\sum_{k=1}^{n}\bar{\lambda}_k\lambda_k(\boldsymbol{x},\ \boldsymbol{e}_k)\,\boldsymbol{e}_k=\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{x}.$$

Thus, equation (5.92) holds for the operators $\boldsymbol{A}$ and $\boldsymbol{A}^*$ and, consequently, the operator $\boldsymbol{A}$ is normal. And this completes the proof of the theorem.

---

* Representation (5.69) is valid for any operator having n pairwise orthogonal eigenvectors.

** We have also used the relations $\boldsymbol{A}\boldsymbol{e}_k=\lambda_k\boldsymbol{e}_k$.

5.8. Canonical Form of Linear Operators

We consider here the choice of a spacial basis for a given linear operator, in which the matrix of the operator has the simplest form called the **Jordan form of a matrix.**

We introduce the concept of a **principal element** of the operator $\boldsymbol{A}$.

Definition. *The element* $\boldsymbol{x}$ *is called the* ***principal element*** *of the operator* $\boldsymbol{A}$ *corresponding to the eigenvalue* λ *if the relations*

$$(\boldsymbol{A}-\lambda\boldsymbol{I})^m\boldsymbol{x}\neq 0,\ (\boldsymbol{A}-\lambda\boldsymbol{I})^{m+1}\boldsymbol{x}=\boldsymbol{0}$$

hold for a certain integral $m\geqslant 1$. *The number* m *here is called the grade of the principal element* x.

In other words, if x is *a principal element of grade* m, *then the element* $(\boldsymbol{A}-\lambda\boldsymbol{I})^m\boldsymbol{x}$ *is an eigenvector of the operator* $\boldsymbol{A}$.

We shall prove here the following *fundamental* theorem.

Theorem 5.32. *Let* $\boldsymbol{A}$ *be a linear operator acting in the n-dimensional Euclidean space* V. *There is a basis*

$$\{\boldsymbol{e}_k^m\},\ k=1,\ 2,\ \ldots,\ l,\ m=1,\ 2,\ \ldots,\ n_k;$$
$$n_1+n_2+\ldots+n_l=n \tag{5.95}$$

which is formed from eigenvectors and principal vectors of the operator $\boldsymbol{A}$ *in which the action of the operator* $\boldsymbol{A}$ *is described by the following relations:*

$$\begin{aligned}&\boldsymbol{A}\boldsymbol{e}_k^1=\lambda_k\boldsymbol{e}_k^1,\ k=1,\ 2,\ \ldots,\ l;\\&\boldsymbol{A}\boldsymbol{e}_k^m=\lambda_k\boldsymbol{e}_k^m+\boldsymbol{e}_k^{m-1},\ k=1,\ 2,\ \ldots,\ l;\ m=2,\ 3,\ \ldots,\ n_k.\end{aligned} \tag{5.96}$$

Before roving the theorem, we shall make some remarks.

Remark 1. The vectors $\boldsymbol{e}_k^1$ $(k=1,\ 2,\ \ldots,\ l)$ of the basis (5.95) are, evidently, the eigenvectors of the operator $\boldsymbol{A}$ which correspond to the eigenvalues λ_k.

From the definitions of principal vectors and relations (5.96) it follows that the vectors $\boldsymbol{e}_k^m$ $(k=1,\ 2,\ \ldots,\ l,\ m=\ 2,\ 3,\ \ldots,\ n_k)$ are principal vectors of grade m corresponding to the eigenvalues λ_k respectively.

Remark 2. Paying attention to formulas (5. 13) and (5.12), we see that relations (5.96) indeed define the action of the operator $\boldsymbol{A}$ in the space V at the given basis $\{\boldsymbol{e}_k^m\}$.

Remark 3. The matrix A of the linear operator $\boldsymbol{A}$ in the basis $\{\boldsymbol{e}_k^m\}$ has the following "block" form:

$$A=\begin{pmatrix}\Lambda_1 & & & 0\\ & \Lambda_2 & & \\ & & \ddots & \\ 0 & & & \Lambda_l\end{pmatrix} \tag{5.97}$$

where the block Λ_k is the following matrix:

$$\Lambda_k = \begin{pmatrix} \lambda_k & 1 & 0 & \dots & 0 \\ 0 & \lambda_k & 1 & & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & & 1 \\ 0 & 0 & 0 & \dots & \lambda_k \end{pmatrix}, \tag{5.98}$$

Remark 4. Form (5.97) of the matrix A of the linear operator $\boldsymbol{A}$ is called the **Jordan form of the matrix** of that operator. The block Λ_k is usually called a **Jordan block** (or submatrix) of the matrix A. Note that Theorem 5.32 on the reduction of the matrix of an operator to the block form (5.97) is known as the theorem on the reduction of the matrix of an operator to the Jordan form.

Remark 5. The Jordan form of matrix (5.97) is defined with an accuracy to within the order of arrangement of the blocks Λ_k along the diagonal of the matrix. This order depends on the order of enumeration of the eigenvalues λ_k.

We present the proof of Theorem 5.32 suggested by A. F. Filippov*.

Proof of Theorem 5.32. We shall prove the theorem by induction. For $n=1$ the assertion of the theorem is obvious. Suppose $n>1$ and the theorem is valid for spaces of dimension smaller than n. We shall prove that under this assumption it is also valid for spaces of dimension n. And this completes the proof of the theorem.

Let λ be an eigenvalue of the operator $\boldsymbol{A}$. In accordance with Theorem 5.8, this number is a root of the characteristic equation $\det(\boldsymbol{A}-\lambda\boldsymbol{I})=0$. Consequently, the rank r of the linear operator**

$$\boldsymbol{B}=\boldsymbol{A}-\lambda\boldsymbol{I} \tag{5.99}$$

is smaller that n, i.e. $r<n$.

The linear operator $\boldsymbol{B}$ maps the space V onto the subspace $\operatorname{im}\boldsymbol{B}$. Therefore, the operator $\boldsymbol{B}$ maps the subspace $\operatorname{im}\boldsymbol{B}$ of dimension $r<n$ into the same subspace. By induction assumption, there is a basis

$$\{\boldsymbol{h}_k^m\},\ k=1,2,\dots,p;\ m=1,2,\dots,r_k;$$
$$r_1+r_2+\dots+r_p=r, \tag{5.100}$$

in $\operatorname{im}\boldsymbol{B}$ in which the action of the operator $\boldsymbol{B}$ from $\operatorname{im}\boldsymbol{B}$ into $\operatorname{im}\boldsymbol{B}$ is given by the following relations:

$$\left.\begin{aligned} &\boldsymbol{B}\boldsymbol{h}_k^1=\mu_k\boldsymbol{h}_k^1,\ k=1,2,\dots,p,\\ &\boldsymbol{B}\boldsymbol{h}_k^m=\mu_k\boldsymbol{h}_k^m+\boldsymbol{h}_k^{m-1},\ k=1,2,\dots,p;\ m=2,3,\dots,r_k. \end{aligned}\right\} \tag{5.101}$$

---

* A. F. Filippov. "A Brief Proof of the Theorem on the Reduction of a Matrix to the Jordan Form". *Vestnik Moscovskogo Universiteta*, 1971, No. 2.

** Recall that the rank r of the linear operator $\boldsymbol{B}$ is equal to the dimension $\operatorname{im}\boldsymbol{B}$, according to Theorem 5.6 the rank r is equal to that of the matrix of the operator.

Thus, in this basis the matrix $\tilde{B}$ of the operator $\tilde{\boldsymbol{B}}$ from $\operatorname{im}\boldsymbol{B}$ into $\operatorname{im}\boldsymbol{B}^*$ has the following block form:

$$\tilde{B}=\begin{pmatrix} M_1 & & & 0 \\ & M_2 & & \\ & & \ddots & \\ 0 & & & M_p \end{pmatrix}, \text{where } M_k=\begin{pmatrix} \mu_k & 1 & 0 & & 0 \\ 0 & \mu_k & 1 & & 0 \\ \cdot & & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & & 1 \\ 0 & 0 & 0 & & \mu_k \end{pmatrix}. \quad (5.102)$$

Suppose that only the first m_1 ($m_1 \geqslant 0$) eigenvalues of the operator $\tilde{B}$ are zero.

Since the rank of each submatrix M_k (see (5.102)), for which $\mu_k=0$, is equal to r_{k-1} and the rank of the submatrix, for which $\mu_k \neq 0$, is equal to r_k, we find, by (5.100), that the rank of the matrix $\tilde{B}$ is equal to $\sum_{k=1}^{p} r_k - m_1 = r - m_1$. Therefore, the dimension of the subspace $\ker\tilde{\boldsymbol{B}}$ is equal to m_1**, and $\ker\tilde{\boldsymbol{B}}$ is a linear closure of the vectors $\boldsymbol{h}_1^1$ and $\boldsymbol{h}_2^1, \ldots, \boldsymbol{h}_{m_1}^1$. By virtue of linear independence, these vectors form a basis in $\ker\tilde{\boldsymbol{B}}$. Evidently, $\ker\tilde{\boldsymbol{B}} \subset \ker\boldsymbol{B}$. Let us complete the basis $\boldsymbol{h}_1^1, \boldsymbol{h}_2^1, \ldots, \boldsymbol{h}_{m_1}^1$ in $\ker\tilde{\boldsymbol{B}}$ by the vectors $\boldsymbol{g}_k$, $k=1, 2, \ldots, m_0$, $m_0=n-r-m_1$ to form a basis in $\ker\boldsymbol{B}$ (by Theorem 5.1, the dimension of $\ker\boldsymbol{B}$ is equal to $n-\dim\operatorname{im}\boldsymbol{B}$, i.e. to $n-r$).

Since $\boldsymbol{g}_k \in \ker\boldsymbol{B}$, we have

$$\boldsymbol{B}\boldsymbol{g}_k=\boldsymbol{0}. \quad (5.103)$$

Let us now consider the vectors $\boldsymbol{h}_k^{r_k}$, $k=1, 2, \ldots, m_1$. Since these vectors belong to $\operatorname{im}\boldsymbol{B}$, there are vectors $\boldsymbol{f}_k \in V$ such that

$$\boldsymbol{B}\boldsymbol{f}_k=\boldsymbol{h}_k^{r_k}, \quad k=1, 2, \ldots, m_1. \quad (5.104)$$

We shall prove that the vectors

$$\left.\begin{array}{l} \boldsymbol{h}_k^m (k=1, 2, \ldots, p; \quad m=1, 2, \ldots, r_k), \\ \boldsymbol{g}_k (k=1, 2, \ldots, m_0), \ \boldsymbol{f}_k (k=1, 2, \ldots, m_1) \end{array}\right\} \quad (5.105)$$

are linearly independent.

Let us consider the linear combination $\boldsymbol{f}$ of these vectors which is equal to zero:

$$\boldsymbol{f}=\sum_{k=1}^{p}\sum_{m=1}^{r_k}\alpha_{km}\boldsymbol{h}_k^m+\sum_{k=1}^{m_0}\beta_k\boldsymbol{g}_k+\sum_{k=1}^{m_1}\gamma_k\boldsymbol{f}_k=\boldsymbol{0}. \quad (5.106)$$

How does the operator $\boldsymbol{B}$ act on the element $\boldsymbol{f}$?

---

* We denote the operator $\boldsymbol{B}$ from $\operatorname{im}\boldsymbol{B}$ into $\operatorname{im}\boldsymbol{B}$ by $\tilde{\boldsymbol{B}}$.

** The rank of the matrix $\tilde{\boldsymbol{B}}$ is equal to $\dim\operatorname{im}\tilde{\boldsymbol{B}}$. According to Theorem 5.1 $\dim\operatorname{im}\tilde{\boldsymbol{B}}+\dim\ker\tilde{\boldsymbol{B}}=r$. Consequently, $\dim\ker\tilde{\boldsymbol{B}}=m_1$.

In accordance with (5.101), (5.103) and (5.104), we get the following expression:

$$\boldsymbol{Bf}=\sum_{k=1}^{p}\alpha_{k1}\mu_k\boldsymbol{h}_k^1+\sum_{k=1}^{p}\sum_{m=2}^{r_k}\alpha_{km}(\mu_k\boldsymbol{h}_k^m+\boldsymbol{h}_k^{m-1})+\sum_{k=1}^{m_1}\gamma_k\boldsymbol{h}_k^{r_k}=\boldsymbol{0}. \qquad (5.107)$$

Relation (5.107) is a linear combination of the base vectors $\{\boldsymbol{h}_k^m\}$ which is equal to zero; therefore, the coefficients in these vectors in the indicated linear combination are zero. Since $\mu_k=0$ for $k\leqslant m_1$, it follows from (5.107) that the coefficients in $\boldsymbol{h}_k^{r_k}$ are exactly equal to γ_k and, therefore, $\gamma_k=0$. From this and from relation (5.106) we get equalities

$$\boldsymbol{g}=\sum_{k=1}^{m_0}\beta_k\boldsymbol{g}_k=-\sum_{k=1}^{p}\sum_{m=1}^{r_k}\alpha_{km}\boldsymbol{h}_k^m, \qquad (5.108)$$

from which it follows that the vector $\boldsymbol{g}$ which is a linear combination of vectors $\{\boldsymbol{g}_k\}$ belongs to $\ker\boldsymbol{B}$ (recall that the vectors $\{\boldsymbol{g}_k\}$ form a part of the basis in $\ker\boldsymbol{B}$).

On the other hand, it follows from (5.108) that $\boldsymbol{g}$ is a linear combination of vectors $\boldsymbol{h}_k^m$, i.e. belongs to $\operatorname{im}\tilde{\boldsymbol{B}}$. Consequently, $\boldsymbol{g}$ belongs to $\ker\tilde{\boldsymbol{B}}$ (recall that $\ker\tilde{\boldsymbol{B}}$ is an intersection of $\operatorname{im}\tilde{\boldsymbol{B}}$ and $\ker\boldsymbol{B}$) and, therefore, $\boldsymbol{g}=\sum_{k=1}^{m_1}\delta_k\boldsymbol{h}_k^1$.

Since the linear closures of the collections of vectors $\{\boldsymbol{g}_k\}$ and $\{\boldsymbol{h}_k^1\}$ only have a zero element in common (taken together, these collections form a basis in $\ker\boldsymbol{B}$) and, as we have established, $\boldsymbol{g}$ belongs to each of the indicated linear closures, we have $\boldsymbol{g}=\boldsymbol{0}$. But then it follows from (5.108) that $\beta_k=0$ $(k=1, 2, \ldots, m_0)$ and $\alpha_{km}=0$ $(k=1, 2, \ldots, p;\ m=1, 2, \ldots, r_k)$.

Thus all the coefficients in the linear combination (5.106) of vectors (5.105) are equal to zero, i.e. vectors (5.105) *are linearly independent.*

The total number of vectors (5.105) is equal to $r+m_0+m_1$. Since $m_0=n-r-m_1$ (we have established this fact above, while proving the theorem, when we introduced the vectors $\boldsymbol{g}_k$), the total number of vectors (5.105) is equal to n and, therefore, they form a basis in V. We introduce the designation

$$\boldsymbol{f}_k=\boldsymbol{h}_k^{r_k+1} \qquad (5.109)$$

and write the vectors forming this basis in the following sequence of

series:

$$\left.\begin{array}{l}\{\boldsymbol{g}_1\};\ \{\boldsymbol{g}_2\};\ \ldots;\ \{\boldsymbol{g}_{m_0}\};\\ \{\boldsymbol{h}_k^1,\ \ldots,\ \boldsymbol{h}_k^{r_k},\ \boldsymbol{h}_k^{r_k+1}\},\quad k=1,\ 2,\ \ldots,\ m_1;\\ \{\boldsymbol{h}_k^1,\ \ldots,\ \boldsymbol{h}_k^{r_k}\},\ k=m_1+1,\ \ldots,\ p.\end{array}\right\}\tag{5.110}$$

Let us see how the operator $\boldsymbol{B}$ acts on the vectors of basis (5.110) in the space V. Considering relations (5.101), (5.103), (5.104) and (5.109), we make sure that the action of $\boldsymbol{B}$ in basis (5.110) is specified by the relations

$$\boldsymbol{B}\boldsymbol{g}_k=0,\ k=1,\ 2,\ \ldots,\ m_0,\ \boldsymbol{B}\boldsymbol{h}_k^{r_k+1}=\boldsymbol{h}_k^{r_k},\ k=1,\ 2,\ \ldots,\ m_1,$$

and relations (5.101).

Thus, in basis (5.110), the operator $\boldsymbol{B}=\boldsymbol{A}-\lambda\boldsymbol{I}$ acts according to rule (5.96), indicated in the formulation of Theorem 5.32. But then, in this basis, the operator $\boldsymbol{A}=\boldsymbol{B}+\lambda\boldsymbol{I}$ also acts by the same rule. We have proved the theorem.

5.9. Linear Operators in a Real Euclidean Space

We shall show here how the results and definitions presented in the previous sections can be extended to the case of real Euclidean spaces.

5.9.1. General remarks. Let us consider an arbitrary n-dimensional real Euclidean space V and an operator $\boldsymbol{A}$ from V into V.

For the case of a real linear space, the concept of a linear operator is formulated by a complete analogy with the respective concept for a complex space.

Definition 1. *The operator $\boldsymbol{A}$ is said to be* ***linear*** *if for any elements $\boldsymbol{x}\in V$ and $\boldsymbol{y}\in V$ and any real numbers α and β the following equality holds true:*

$$\boldsymbol{A}(\alpha\boldsymbol{x}+\beta\boldsymbol{y})=\alpha\boldsymbol{A}\boldsymbol{x}+\beta\boldsymbol{A}\boldsymbol{y}.\tag{5.111}$$

The concepts of an eigenvalue and an eigenvector of an operator are introduced by a complete analogy with a complex space.

It should be pointed out that eigenvalues are roots of the characteristic equation of an operator.

The converse statement in a real case is valid only when the corresponding root of the characteristic equation is real. Only in that case is the indicated root an eigenvalue of the linear operator in question.

In this connection, it is natural to isolate some class of linear operators in a real Euclidean space, all roots of whose characteristic equations are real.

We have proved in Theorem 5.16 that all the eigenvalues of a self-adjoint operator are real. In addition, the concept of a self-adjoint

operator played an important part in the inferences concerning quadratic forms made in 5.6. It is, therefore, natural to extend the notion of a self-adjoint operator to the case of a real space.

We shall first introduce the concept of the operator $\boldsymbol{A}^*$ which is an adjoint of the operator $\boldsymbol{A}$. Namely, the operator $\boldsymbol{A}^*$ is said to be an adjoint of $\boldsymbol{A}$ if for any $\boldsymbol{x}$ and $\boldsymbol{y}$ from V the equality $(\boldsymbol{Ax}, \boldsymbol{y})=(\boldsymbol{x}, \boldsymbol{A}^*\boldsymbol{y})$ is satisfied.

Without any difficulty we can extend Theorem 5.12 on the existence and uniqueness of an adjoint operator to the case of a real space.

Recall that the proof of Theorem 5.12 is based on the concept of a sesquilinear form. In the real case, instead of a sesquilinear form use should be made of the bilinear form $B(\boldsymbol{x}, \boldsymbol{y})$. A requisite remark was made in this connection in 5.4.2.

Let us recall, in this connection, the definition of a bilinear form in any real, not necessarily Euclidean, linear space L. Let B be a function putting every ordered pair $(\boldsymbol{x}, \boldsymbol{y})$ of vectors $\boldsymbol{x}\in L$ and $\boldsymbol{y}\in L$ into correspondence with the real number $B(\boldsymbol{x}, \boldsymbol{y})$.

Definition 2. *The function $B(\boldsymbol{x}, \boldsymbol{y})$ is called a* ***bilinear form*** *given on L if for any vectors $\boldsymbol{x}, \boldsymbol{y}$ and $\boldsymbol{z}$ from L and for any real number λ the relations*

$$\left.\begin{aligned} B(\boldsymbol{x}+\boldsymbol{z}, \boldsymbol{y})&=B(\boldsymbol{x}, \boldsymbol{y})+B(\boldsymbol{z}, \boldsymbol{y}),\\ B(\boldsymbol{x}, \boldsymbol{y}+\boldsymbol{z})&=B(\boldsymbol{x}, \boldsymbol{y})+B(\boldsymbol{x}, \boldsymbol{z}),\\ B(\lambda\boldsymbol{x}, \boldsymbol{y})&=B(\boldsymbol{x}, \lambda\boldsymbol{y})=\lambda B(\boldsymbol{x}, \boldsymbol{y}),\end{aligned}\right\} \tag{5.112}$$

hold true.

An important part will be played in this section by the special representation of the bilinear form $B(\boldsymbol{x}, \boldsymbol{y})$ as

$$B(\boldsymbol{x}, \boldsymbol{y})=(\boldsymbol{Ax}, \boldsymbol{y}), \tag{5.113}$$

where $\boldsymbol{A}$ is some linear operator. The corresponding theorem (Theorem 5.11) on an analogous representation of a sesquilinear form in a complex space was based on the inferences made in the lemma in 5.4.1 on a special representation of the linear form $f(\boldsymbol{x})$. It was indicated at the end of 5.4.1 that the lemma was true in a real space as well. It should only be pointed out that in the proof of the lemma the choice of the elements $\boldsymbol{h}^k$ must be performed not by formula (5.41) but by the formula $\boldsymbol{h}^k=f(\boldsymbol{e}_k)$, where $f(\boldsymbol{x})$ is a given linear form in a real space.

In 5.6 we introduced Hermitian forms. The Hermitian form is the sesquilinear form $B(\boldsymbol{x}, \boldsymbol{y})$ in a complex space characterized by the relation $B(\boldsymbol{x}, \boldsymbol{y})=\overline{B(\boldsymbol{y}, \boldsymbol{x})}$ (the bar over B signifies that we take a complex conjugate value of B).

In the case of a real space, the analogs of Hermitian forms are symmetric bilinear forms. Such a form is characterized by the relation

$$B(\boldsymbol{x}, \boldsymbol{y})=B(\boldsymbol{y}, \boldsymbol{x}). \tag{5.114}$$

The bilinear form $B(\boldsymbol{x}, \boldsymbol{y})$ given in the linear space L is said to be skew-symmetric if the relation $B(\boldsymbol{x}, \boldsymbol{y}) = -B(\boldsymbol{y}, \boldsymbol{x})$ holds true for any vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ belonging to L. It is evident that for every bilinear form the functions

$$B_1(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}[B(\boldsymbol{x}, \boldsymbol{y}) + B(\boldsymbol{y}, \boldsymbol{x})],$$

$$B_2(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}[B(\boldsymbol{x}, \boldsymbol{y}) - B(\boldsymbol{y}, \boldsymbol{x})]$$

are a symmetric and a skew-symmetric form respectively. Since $B(\boldsymbol{x}, \boldsymbol{y}) = B_1(\boldsymbol{x}, \boldsymbol{y}) + B_2(\boldsymbol{x}, \boldsymbol{y})$, we obtain the following statement: *Any bilinear form can be represented as the sum of symmetric and skew-symmetric bilinear forms.*

It is easy to see that such a representation is unique.

We shall prove the following theorem on symmetric bilinear forms (this theorem is an analog of Theorem 5.25 on Hermitian forms).

Theorem 5.33. *For the bilinear form $B(\boldsymbol{x}, \boldsymbol{y})$ specified on various vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ of the real Euclidean space V to be symmetric, it is necessary and sufficient that the linear operator $\boldsymbol{A}$ appearing in representation* (5.113) *should be self-adjoint.*

Proof. If $\boldsymbol{A}$ is a self-adjoint operator, we can use the properties of a scalar product and get

$$B(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}, \boldsymbol{A}\boldsymbol{y}) = (\boldsymbol{A}\boldsymbol{y}, \boldsymbol{x}) = B(\boldsymbol{y}, \boldsymbol{x}).$$

Thus relation (5.114) holds true, i.e. the bilinear form $B(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{A}\boldsymbol{x}, \boldsymbol{y})$ is symmetric.

Now if the form $B(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{A}\boldsymbol{x}, \boldsymbol{y})$ is symmetric, then the relations $(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}) = B(\boldsymbol{x}, \boldsymbol{y}) = B(\boldsymbol{y}, \boldsymbol{x}) = (\boldsymbol{A}\boldsymbol{y}, \boldsymbol{x})$ hold true.

Consequently, the operator $\boldsymbol{A}$ is self-adjoint. And this is what we wished to prove.

We introduce the concept of the matrix of the linear operator $\boldsymbol{A}$. Let $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ be some basis in the n-dimensional real linear space L. We set $\boldsymbol{A}\boldsymbol{e}_k = \sum_{i=1}^{n} a_k^i \boldsymbol{e}_i$.

Then, as in a complex case, it is easy to show that if $\boldsymbol{x} = \sum_{k=1}^{n} x^k \boldsymbol{e}_k$, then the representation $y^i = \sum_{k=1}^{n} a_k^i x^k$ holds for the components of the vector $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$.

The matrix $A = (a_k^i)$ is known as the matrix of the linear operator $\boldsymbol{A}$ in the basis $\{\boldsymbol{e}_k\}$.

By analogy with what was done in 5.2, we can prove that the quantity $\det A$ does not depend on the choice of the basis, and thus we correctly define $\det \boldsymbol{A}$ of the operator $\boldsymbol{A}$.

The *characteristic equation* corresponding to the operator $\boldsymbol{A}$ is an equation $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$, and the polynomial appearing on the left-

hand side of this equation is called a *characteristic polynomial of the operator* $\boldsymbol{A}$.

Let us now prove the theorem on the roots of a characteristic polynomial of a self-adjoint operator in a real Euclidean space.

Theorem 5.34. *All the roots of the characteristic polynomial of the self-adjoint linear operator* $\boldsymbol{A}$ *in a Euclidean space are real.*

Proof. Suppose $\lambda=\alpha+i\beta$ is a root of the characteristic equation

$$\det(\boldsymbol{A}-\lambda\boldsymbol{I})=0 \tag{5.115}$$

of the self-adjoint operator $\boldsymbol{A}$.

We fix some basis $\{\boldsymbol{e}_k\}$ in V and designate as a_{jk} the elements of the matrix of the operator $\boldsymbol{A}$ in that basis (note that a_{jk} are real numbers).

We shall seek a nonzero solution of the following system of homogeneous linear equations with respect to $\xi_1, \xi_2, \ldots, \xi_n$:

$$\sum_{k=1}^{n} a_{jk}\xi_k=\lambda\xi_j, \quad j=1, 2, \ldots, n, \tag{5.116}$$

where $\lambda=\alpha+i\beta$.

Since the determinant of system (5.116) is equal to $\det(\boldsymbol{A}-\lambda\boldsymbol{I})$ (recall that the determinant of the matrix of a linear transformation does not depend on the choice of the basis and, according to (5.115), that determinant is zero), system (5.116) of the homogeneous linear equations has a nonzero solution $\xi_k=x_k+iy_k$, $k=1, 2, \ldots, n$.

Substituting this solution into the right-hand and left-hand sides of system (5.116), taking into account that $\lambda=\alpha+i\beta$ and separating the real and the imaginary part of the relations obtained, we find that the collections $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$ of real numbers* satisfy the following system of equations:

$$\left.\begin{aligned} &\sum_{k=1}^{n} a_{jk}x_k=\alpha x_j-\beta y_j, \\ &\sum_{k=1}^{n} a_{jk}y_k=\alpha y_j+\beta x_j, \quad j=1, 2, \ldots, n. \end{aligned}\right\} \tag{5.117}$$

Let us consider the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ with the coordinates $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$, respectively, in the given basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$. Then we can rewrite relations (5.117) in the form

$$\boldsymbol{A}\boldsymbol{x}=\alpha\boldsymbol{x}-\beta\boldsymbol{y}, \quad \boldsymbol{A}\boldsymbol{y}=\alpha\boldsymbol{y}+\beta\boldsymbol{x}.$$

Multiplying scalarly the first of the relations obtained by $\boldsymbol{y}$ and the second relation by $\boldsymbol{x}$, we, evidently, obtain equalities

$$\left.\begin{aligned} &(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{y})=\alpha(\boldsymbol{x}, \boldsymbol{y})-\beta(\boldsymbol{y}, \boldsymbol{y}), \\ &(\boldsymbol{x}, \boldsymbol{A}\boldsymbol{y})=\alpha(\boldsymbol{x}, \boldsymbol{y})+\beta(\boldsymbol{x}, \boldsymbol{x}). \end{aligned}\right\} \tag{5.118}$$

---

* Recall that not all these numbers are zero.

The operator $\boldsymbol{A}$ being self-adjoint, we have $(\boldsymbol{Ax}, \boldsymbol{y})=(\boldsymbol{x}, \boldsymbol{Ay})$. Therefore, subtracting relations (5.118), we get an equation

$$\beta[(\boldsymbol{x}, \boldsymbol{x})+(\boldsymbol{y}, \boldsymbol{y})]=0.$$

But $(\boldsymbol{x}, \boldsymbol{x})+(\boldsymbol{y}, \boldsymbol{y})\neq 0$ (if $(\boldsymbol{x}, \boldsymbol{x})+(\boldsymbol{y}, \boldsymbol{y})=0$, then $x_k=0$ and $y_k=0$, $k=1, 2, \ldots, n$; consequently, the solution $\xi_k=x_k+iy_k$ would be zero whereas, by construction, this solution is nonzero). Therefore, $\beta=0$, and since β is an imaginary part of the root $\lambda=\alpha+i\beta$ of the characteristic equation (5.115), λ is, evidently, a real number. We have proved the theorem.

As in the complex case, the statement concerning the existence of an orthonormal basis consisting of the eigenvectors of that operator holds true for a self-adjoint operator (the analog of theorem 5.21). Let us prove this statement.

Theorem 5.35. *Every self-adjoint linear operator $\boldsymbol{A}$ acting in the n-dimensional real Euclidean space V possesses an orthonormal basis of eigenvectors.*

Proof. Let λ_1 be a real eigenvalue of the operator $\boldsymbol{A}$, and $\boldsymbol{e}_1$ is a unit eigenvector corresponding to that eigenvalue ($\| \boldsymbol{e}_1 \| = 1$).

We designate as V_1 an $(n-1)$-dimensional subspace of the space V which is orthogonal to $\boldsymbol{e}_1$. V_1 is evidently an invariant subspace of the space V (i.e. if $\boldsymbol{x}\in V_1$, then $\boldsymbol{Ax}\in V_1$). Indeed, suppose $\boldsymbol{x}\in V_1$; then $(\boldsymbol{x}, \boldsymbol{e}_1)=0$. Since the operator $\boldsymbol{A}$ is self-adjoint and λ_1 is an eigenvalue of $\boldsymbol{A}$, we get $(\boldsymbol{Ax}, \boldsymbol{e}_1)=(\boldsymbol{x}, \boldsymbol{Ae}_1)=\lambda_1(\boldsymbol{x}, \boldsymbol{e}_1)=0$.

Consequently, $\boldsymbol{Ax}\in V_1$ and, therefore, V_1 is an invariant subspace of the operator $\boldsymbol{A}$. We can, therefore, consider the operator $\boldsymbol{A}$ in the subspace V_1. It is clear that in V_1 the operator $\boldsymbol{A}$ is self-adjoint. By Theorem 5.34, the operator $\boldsymbol{A}$ acting in V_1 has a real eigenvalue λ_2 which is associated with an eigenvector $\boldsymbol{e}_2\in V_1$ of the operator $\boldsymbol{A}$ satisfying the condition $\| \boldsymbol{e}_2 \| = 1$.

Considering then an $(n-2)$-dimensional subspace of V_2 which is orthogonal to the vectors $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ and repeating the arguments we have just presented, we construct an eigenvector $\boldsymbol{e}_3$ of the operator $\boldsymbol{A}$ which is orthogonal to the vectors $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ and satisfies the condition $\| \boldsymbol{e}_3 \| = 1$.

Continuing with the same reasoning, we find, as a result, n mutually orthogonal eigenvectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ of the operator $\boldsymbol{A}$ satisfying the condition $\| \boldsymbol{e}_k \| = 1$, $k=1, 2, \ldots, n$. The vectors $\{\boldsymbol{e}_k\}$ evidently form a basis in V. And this is what we set out to prove.

Remark. Assume that $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ is an orthonormal basis in the n-dimensional Euclidean space V consisting of the eigenvectors of the self-adjoint operator $\boldsymbol{A}$, i.e. $\boldsymbol{Ae}_k=\lambda_k\boldsymbol{e}_k$. Then the matrix of the operator $\boldsymbol{A}$ in the basis $\{\boldsymbol{e}_k\}$ is diagonal and the diagonal elements have the form $a_k^k=\lambda_k$.

Note that if $\{\boldsymbol{e}_k\}$ is an arbitrary orthonormal basis in the real Euclidean space V, then the matrix of the self-adjoint operator $\boldsymbol{A}$ is sym-

metric, i.e. $A' = A$. The converse statement is also true, i.e. if in a certain orthonormal basis $\{ee_k\}$ the matrix of an operator is symmetric, then the operator A is self-adjoint.

This is what distinguishes a real case from a complex one, since in a complex case the operator A is self-adjoint if and only if the matrix A of that operator is Hermitian in an orthonormal basis, i.e. the elements a_k^i of the matrix A satisfy the condition $a_k^i = \bar{a}_i^k$ (the bar signifies complex conjugation).

The indicated assertion follows directly from the fact that if (a_k^i) is a matrix of the operator A, then the matrix of the conjugate operator is equal to (a_i^k) in a real case and to $(\bar{a}_i^k)$ in a complex case, which can be easily verified by direct calculations.

5.9.2. Orthogonal operators. In a complex Euclidean space an important role is played by unitary operators introduced in 5.7. Orthogonal operators are the analog of unitary operators in a real Euclidean space.

Definition 1. *The linear operator P acting in the real Euclidean space V is said to be* ***orthogonal*** *if for any x and y belonging to V the equality*

$$(Px, Py) = (x, y) \tag{5.119}$$

holds true.

Thus, an orthogonal operator retains a scalar product. Hence it follows directly that if $e_1, e_2, \ldots, e_n$ is an orthonormal basis of the Euclidean space V, then $Pe_1, Pe_2, \ldots, Pe_n$ is also an orthonormal basis. In what follows, we shall call condition (5.119) the condition of orthogonality of the operator P.

The following statement holds true.

Theorem 5.36. *For the linear operator P to be orthogonal, it is necessary and sufficient that there exist an operator P^{-1} and there hold an equation*

$$P^* = P^{-1}, \tag{5.120}$$

where P^* is an adjoint operator of P, and *P^{-1} is an operator inverse with respect to P.*

Proof. **(1) Necessity.** Let P be an orthogonal operator, i.e. condition (5.119) be satisfied. Applying the adjoint operator P^*, we can rewrite this condition as $(P^*Px, y) = (x, y)$.

Thus, for any x and y, the equality $((P^*P - I)x, y) = 0$ holds true.

Fixing any element x in this equality and considering y to be arbitrary, we find that the linear operator $P^*P - I$ acts according to the rule $(P^*P - I)x = 0$.

Consequently, $P^*P = I$; we can verify by a complete analogy that $PP^* = I$. Thus the operators P^* and P are inverse of each other, i.e. condition (5.120) is fulfilled.

(2) Sufficiency. Assume that condition (5.120) is fulfilled. Then, evidently, $\boldsymbol{PP^*=P^*P=I}$.

Directing our attention to an adjoint operator and using the relations we have just written, we get, for any $\boldsymbol{x}$ and $\boldsymbol{y}$, relations $(\boldsymbol{Px}, \boldsymbol{Py})=(\boldsymbol{x}, \boldsymbol{P^*Py})=(\boldsymbol{x}, \boldsymbol{Iy})=(\boldsymbol{x}, \boldsymbol{y})$.

We see that condition (5.119) is fulfilled and, consequently, the operator $\boldsymbol{P}$ is orthogonal. We have thus proved the theorem.

We shall introduce now the concept of an orthogonal matrix P.

Definition 2. *The matrix* $\boldsymbol{P}$ *is said to be orthogonal if*

$$P'P=PP'=I, \tag{5.121}$$

where P' *is a transposed matrix and* I *is an identity matrix.*

If $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ is an orthonormal basis in the Euclidean space V, then the operator $\boldsymbol{P}$ is orthogonal if and only if its matrix in the basis $\{\boldsymbol{e}_k\}$ is orthogonal.

It follows directly from equation (5.121) that if the matrix $P=(p_i^k)$ is orthogonal then

$$\sum_{i=1}^{n} p_i^k p_i^l = \begin{cases} 1 & \text{for } k=l, \\ 0 & \text{for } k\neq l. \end{cases}$$

In a complex Euclidean space a *unitary* matrix is an analog of an orthogonal matrix. Namely, the matrix U is said to be *unitary* if there holds a relation

$$U^*U=UU^*=I, \tag{5.122}$$

in which U^* is a Hermitian conjugate matrix, i.e. $U^*=\bar{U}'$, where the prime means a transposition and the bar, a complex conjugation.

It is easy to show that in an orthonormal basis the matrix of the linear operator $\boldsymbol{U}$ is unitary if and only if the operator $\boldsymbol{U}$ is unitary.

In conclusion we shall consider an example of an orthogonal transformation in one-dimensional and two-dimensional spaces.

In a one-dimensional case, every vector $\boldsymbol{x}$ has the form $\boldsymbol{x}=\alpha\boldsymbol{e}$, where α is a real number and $\boldsymbol{e}$ is a vector generating the given space. Then $\boldsymbol{Pe}=\lambda\boldsymbol{e}$, and since $(\boldsymbol{Pe}, \boldsymbol{Pe})=\lambda^2(\boldsymbol{e}, \boldsymbol{e})=(\boldsymbol{e}, \boldsymbol{e})$, we have $\lambda=\pm 1$.

Thus, in a one-dimensional case there are two orthogonal transformations, $\boldsymbol{P}_+\boldsymbol{x}=\boldsymbol{x}$ and $\boldsymbol{P}_-\boldsymbol{x}=-\boldsymbol{x}$.

In a two-dimensional case, every orthogonal transformation is defined in an arbitrary orthonormal basis by an orthogonal matrix of order 2, i.e. the matrix $P=\begin{pmatrix} a & b \\ c & d \end{pmatrix}$. From the condition $\boldsymbol{PP'}=\boldsymbol{P'P}=\boldsymbol{I}$ it follows that

$$a^2+b^2=1,\ a^2=d^2,\ b^2=c^2,\ ac+db=0,\ ab+cd=0.$$

Setting $a=\cos\varphi$, $b=-\sin\varphi$, we find that every orthogonal matrix

order 2 has the form

$$P_{\pm}=\begin{pmatrix} \cos\varphi & -\sin\varphi \\ \pm\sin\varphi & \pm\cos\varphi \end{pmatrix}.$$

In both cases, we must take either the plus sign or the minus sign in the second row.

Note that $\det P_{\pm}=\pm 1$. The orthogonal matrix P_+ is said to be *proper* and the orthogonal matrix P_-, *improper*.

The operator $\boldsymbol{P}_+$ with the matrix P_+ performs, in the orthonormal basis $\boldsymbol{e}_1$, $\boldsymbol{e}_2$, a rotation in the plane $\boldsymbol{e}_1$, $\boldsymbol{e}_2$ through the angle φ.

To find out how the operator $\boldsymbol{P}_-$ with the matrix P_- acts, we introduce the matrix $Q=\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ coinciding with P_- for $\varphi=0$, and note that $P_-=QP_+$. The matrix Q is associated with the reflection of the plane with respect to the $\boldsymbol{e}_1$ axis and, consequently, the action of the operator $\boldsymbol{P}_-$ consists in a rotation through the angle φ and in subsequent reflection.

Note that the vectors $\boldsymbol{P}_+\boldsymbol{e}_1$, $\boldsymbol{P}_+\boldsymbol{e}_2$ form an orthonormal basis by virtue of the orthogonality of P_+ and in that basis the matrix of the operator $\boldsymbol{P}_-$ coincides with Q, i.e. is diagonal.

In a general case, when the orthogonal operator $\boldsymbol{P}$ acts in an n-dimensional Euclidean space, there is an orthonormal basis $\boldsymbol{e}_1$, $\boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ in which the matrix of the operator $\boldsymbol{P}$ has the form

$$\begin{pmatrix}
1 & & & & & & & & & & \\
 & 1 & & & & & & & & & \\
 & & \ddots & & & & & & & 0 & \\
 & & & -1 & & & & & & & \\
 & & & & -1 & & & & & & \\
 & & & & & \cos\varphi_1 & -\sin\varphi_1 & & & & \\
 & & & & & \sin\varphi_1 & -\cos\varphi_1 & & & & \\
 & & & & & & & \ddots & & & \\
 & & & & & & & & \ddots & & \\
 & 0 & & & & & & & & \cos\varphi_k & -\sin\varphi_k \\
 & & & & & & & & & \sin\varphi_k & -\cos\varphi_k
\end{pmatrix}$$

In this matrix, all the elements, except for those written out, are equal to zero.

Thus, in a certain orthonormal basis, the action of an orthogonal operator reduces to successive rotations and reflections with respect to the coordinate axes.

Chapter 6

ITERATIVE METHODS OF SOLVING LINEAR SYSTEMS AND EIGENVALUE PROBLEMS

In this chapter we study various methods of solving systems of linear equations with real coefficients with respect to the unknowns which also assume real values.

All methods of solving systems of linear equations which find practical applications can be divided into two large groups, **exact** methods and **iterative** methods.

An **exact** method is a method making it theoretically possible to obtain exact values of the unknowns as a result of a **finite** number of arithmetical operations. Cramer's rule*, presented in Chapter 3, can serve as an example of an exact method.

Iterative methods enable us to find a solution only as a limit of a sequence of vectors which are constructed by means of a uniquely defined process called a process of iterations (successive approximations). Iterative methods are very convenient for the use of modern computers.

The first section of the present chapter is devoted to the exposition of the most applicable iterative methods of solving linear systems.

Iterative methods also find wide application in solving another very important computing problem of linear algebra, the so-called **complete problem of eigenvalues** (this is the name used for the problem of seeking all eigenvalues and the corresponding eigenvectors of a given matrix**). In iterative methods the eigenvalues are calculated as the limits of some numerical sequences without first determining the coefficients in a characteristic polynomial.

In 6.2 we shall discuss one of the most important methods of solving a complete problem of eigenvalues used in computing tech-

---

* The method based on Cramer's rule is seldom used in practice since it requires a large number of arithmetical operations and notations. The exact method based on successive elimination of the unknowns and known as the **Gaussian method** is more convenient.

** As distinct from this problem, the problem of finding some (say, the largest in absolute value) eigenvalues of a given matrix is known as a **partial problem of eigenvalues.**

nique, the so-called **rotation method** (or **Jacobi's method** for computing eigenvalues). This method can be applied to any symmetric (or Hermitian) matrix, can be easily realized on computers and always converges. It is stable as far as the errors of rounding-off the results of intermediary calculations are concerned and possesses a remarkable property that the existence of multiple eigenvalues, close to one another, not only does not retard its convergence but, on the contrary, accelerates it. Jacobi's method known from the middle of the nineteenth century did not find practical application for a long time because of a large number of calculations necessary for its realization. Only the advent of fast electronic computers turned it into the most efficient method of solving the complete problem of eigenvalues of symmetric and Hermitian matrices.

6.1. Iterative Methods of Solving Linear Systems

6.1.1. The method of simple iteration (Jacobi's method). Let us consider a quadratic system of linear equations with real coefficients (3.10) (see 3.2.1), which we shall write in matrix form

$$AX=F, \tag{6.1}$$

assuming A to be the matrix of the coefficients of the linear equations

$$A=\begin{Vmatrix} a_{11} & a_{12} & & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{2n} & & a_{nn} \end{Vmatrix}, \tag{6.2}$$

and X and F to be the vector-columns of the form

$$X=\begin{Vmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{Vmatrix}, \quad F=\begin{Vmatrix} f_1 \\ f_2 \\ \cdots \\ f_n \end{Vmatrix},$$

the first of which must be defined and the second one is given.

Assuming a single-valued solvability of system (6.1), we replace the matrix equation (6.1) by the equivalent matrix equation $X=X-\tau AX+\tau F$ in which τ denotes a real number known as a **stationary parameter.**

With the aid of the last equation we form an iterative sequence of the vectors $\{X_k\}$ defining it by the recurrence relation

$$X_{k+1}=X_k-\tau AX_k+\tau F \quad (k=0,\ 1,\ \ldots) \tag{6.3}$$

with an arbitrary choice of the "zero" approximation X_0.

The method of simple iteration consists in replacing an exact solution X of system (6.1) by the kth iteration X_k with a sufficiently large number k. Let us estimate the **error** $Z_k = X_k - X$ of the method of simple iteration.

Relations (6.3) and (6.1) immediately yield the following matrix equation for the error Z_k:

$$Z_{k+1} = (I - \tau A) Z_k, \tag{6.4}$$

where I is an identity matrix of order n.

Let us consider the norm of a vector in the space E^n and the operator form of a square matrix of order n. As is customary, we use the term the **norm of the vector** X for the number $\| X \|$ which is equal to the square root of the sum of the squares of the coordinates of that vector. The **operator norm** of the abitrary **matrix** A is the number $\| A \|$ equal to either the least upper bound of the relation $\| AX \| / \| X \|$ on the set of all nonzero vectors X or (which is the same thing) the greatest lower bound of the norms $\| AX \|$ on the set of all vectors X having a norm equal to unity.

Thus, by definition,

$$\| A \| = \sup_{X \neq 0} \frac{\| AX \|}{\| X \|}. \tag{6.5}$$

Recall that for any symmetric matrix A* the operator norm of that matrix is equal to the eigenvalue of this matrix with the greatest absolute value (see 5.5.4), i.e.

$$\| A \| = \max_s | \lambda_s |. \tag{6.6}$$

Equation (6.5) yields the following inequality valid for any matrix A and any vector X:

$$\| AX \| \leqslant \| A \| \, \| X \|. \tag{6.7}$$

From the matrix equation for an error (6.4) and from inequality (6.7) we find that for any number k

$$\| Z_{k+1} \| \leqslant \| I - \tau A \| \, \| Z_k \|. \tag{6.8}$$

Let us now prove the following simple but important theorem.

Theorem 6.1. *For the iterative sequence* (6.3) *to converge to the exact solution X of system* (6.1) *for any choice of the zero approximation X_0 and for the given value of the parameter τ, it is sufficient that the condition*

---

* The matrix A is said to be symmetric if $A = A'$.

$$\rho = \| I - \tau A \| < 1 \tag{6.9}$$

be satisfied. In this case, sequence (6.3) *converges with the speed of a geometric progression with the common ratio* ρ.

When the matrix A is symmetric, condition (6.9) *is also a necessary condition for the convergence of the iterative sequence* (6.3) *for any choice of the zero approximation* X_0.

Proof. To establish the sufficiency of condition (6.9), we note that inequality (6.8) yields the following relation:

$$\| Z_k \| \leqslant \| I - \tau A \|^k \| Z_0 \|. \tag{6.10}$$

It is evident from (6.10) that condition (6.9) ensures the convergence of the sequence of errors Z_k to zero with the speed of a geometric progression with the common ratio ρ.

When the matrix A is symmetric, the matrix $I - \tau A$ is also symmetric and, therefore, by virtue of (6.6), condition (6.9) can be rewritten in the equivalent form

$$\rho = \max_s \; | 1 - \tau \lambda_s | < 1 \tag{6.11}$$

(the symbol $\{\lambda_s\}$ denotes here the eigenvalues of the matrix A).

Let us verify the fact that condition (6.11) is the necessary condition for the convergence of the sequence $\{Z_k\}$ to zero for any choice of the zero approximation X_0. We assume that condition (6.11) is not satisfied. Then there is an eigenvalue λ_s satisfying inequality $| 1 - \tau \lambda_s | \geqslant 1$. We denote by $X^{(s)}$ the eigenvector of the matrix A corresponding to that eigenvalue and choose a zero approximation X_0 such that Z_0 coincides with $X^{(s)}$. Then, writing successively relation (6.4) for the numbers $1, 2, \ldots, k$, we find that $Z_k = (1 - \tau \lambda_s)^k Z_0$. By virtue of the inequality $| 1 - \tau \lambda_s | \geqslant 1$, it follows from the last relation that $\| Z_k \|$ does not tend to zero as $k \to \infty$. We have proved Theorem 6.1.

We draw attention to the fact that for practical applications it is not enough to establish the fact of the convergence of the sequence of iterations. The central problem of numerical methods is the estimation of the speed of convergence. It is very important to know the best way to use the stationary parameter τ to obtain the fastest convergence. Let us discuss this problem in more detail.

Suppose we are given an ε-accuracy with which we must obtain an exact solution of system (6.1). It is required to find an iteration X_k with a number k such that

$$\| Z_k \| \leqslant \varepsilon \| Z_0 \|. \tag{6.12}$$

It follows from (6.9) and (6.10) that $\| Z_k \| \leqslant \rho^k \| Z_0 \|$ and, therefore, (6.12) holds true for $\rho^k \leqslant \varepsilon$, i.e. for $k \geqslant \frac{\ln(1/\varepsilon)}{\ln(1/\rho)}$.

This shows that to decrease the number of iterations k sufficient to achieve the required ε-accuracy, we must choose the parameter τ such that the minimum of the function $\rho=\rho(\tau)=\| I-\tau A \|$ is attained.

Assuming the matrix A to be *symmetric* and positive definite, we arrive at the following optimization problem: it is required to find the minimum of the function

$$\min_{\tau} \rho(\tau)=\min_{\tau} \| I-\tau A \|=\min_{\tau} \{\max_{s} | 1-\tau\lambda_s |\}.$$

The solution of this and of a somewhat more general problem suggested by A.A. Samarsky is presented in 6.1.2. We shall prove there that the indicated minimum of the function $\rho=\rho(\tau)$ can be attained for the value $\tau=2/(\gamma_1+\gamma_2)$, where γ_1 and γ_2 are the minimum and the maximum eigenvalues of the matrix A respectively, the minimum value of the function $\rho(\tau)$ being equal to

$$\frac{1-\frac{\gamma_1}{\gamma_2}}{1+\frac{\gamma_1}{\gamma_2}}=\frac{\gamma_2-\gamma_1}{\gamma_2+\gamma_1}.$$

6.1.2. The general implicit method of simple iteration. We shall again consider the solution of linear system (6.1), but this time we replace the iterative sequence (6.3) by a more general iterative sequence defined by the relation

$$B\frac{X_{k+1}-X_k}{\tau}+AX_k=F, \tag{6.13}$$

in which B is some "easily invertible" square matrix of order n and τ is a stationary parameter. This method of forming an iterative sequence is called an **implicit method of simple iteration.** The explicit method of simple iteration considered in 6.1.1 can be obtained from the implicit method in the special case $B=I$, where I is an identity matrix of order n.

To formulate the condition of convergence of the general implicit method of simple iteration in a form convenient for application, we must recall certain notions introduced in the previous chapter.

Recall that the matrix A is called **positive definite** if $(AX, X)>0$ for any nonzero vector X. It was proved in Chapter 5 that the necessary and sufficient condition for the positive definiteness of the symmetric matrix A (or, what is the same, of a self-adjoint linear operator $\boldsymbol{A}$) was the positiveness of all the eigenvalues of that matrix (that operator).

If the matrix A is positive definite, then we agree to write the inequality $A>0$. Next we agree to write the inequality $B>A$ (or $A<B$) when $B-A>0$ (i.e. when the matrix $B-A$ is positive definite).

We shall prove the following remarkable theorem*.

Theorem 6.2 (Samarsky's theorem). *Let the matrix A be symmetric and the conditions $A>0$, $B>0$ be fulfilled (symmetry of the matrix B is, in general, not assumed).*

Then, for the iterative sequence defined by the relations

$$B\frac{X_{k+1}-X_k}{\tau}+AX_k=F \tag{6.13}$$

to converge to the exact solution X of the system $AX=F$ for any choice of the zero approximation X_0, it is sufficient that the following conditions should be satisfied:

$$2B>\tau A,\ \tau A>0. \tag{6.14}$$

Under an additional assumption that the matrix B is symmetric, conditions (6.14) *are not only sufficient but also necessary for the convergence of the indicated iterative sequence for any choice of the zero approximation X_0.*

Proof. (1) **Sufficiency.** Let us first estimate the error $Z_k=X_k-X$. Since X satisfies the equation $AX=F$ and X_k satisfies relation (6.13), we get, for Z_k, a relation

$$B\frac{Z_{k+1}-Z_k}{\tau}+AZ_k=0. \tag{6.15}$$

Let us establish for the error Z_k the so-called basic **energy relation.**

Multiplying (6.15) scalarly by the vector

$$2(Z_{k+1}-Z_k)=2\tau\frac{Z_{k+1}-Z_k}{\tau},$$

we obtain an equation

$$2\tau\left(B\frac{Z_{k+1}-Z_k}{\tau},\ \frac{Z_{k+1}-Z_k}{\tau}\right)+2\tau\left(AZ_k,\frac{Z_{k+1}-Z_k}{\tau}\right)=0. \tag{6.16}$$

If we use the designation $C=2B-\tau A$ and the relation

$$Z_k=\frac{Z_{k+1}+Z_k}{2}-\frac{Z_{k+1}-Z_k}{2}=\frac{Z_{k+1}-Z_k}{2}-\frac{\tau}{2}\cdot\frac{Z_{k+1}-Z_k}{\tau},$$

we can rewrite equation (6.16) in the form

$$\tau\left(C\frac{Z_{k+1}-Z_k}{\tau},\ \frac{Z_{k+1}-Z_k}{\tau}\right)+(A(Z_{k+1}+Z_k),\ Z_{k+1}-Z_k)=0. \tag{6.17}$$

* This theorem is a special case of a considerably more general statement proved by the eminent Soviet mathematician A. A. Samarsky.

Next we note that the matrix A being symmetric, the second term in (6.17) is equal to $(AZ_{k+1}, Z_{k+1})-(AZ_k, Z_k)$. This leads us to the basic **energy relation**

$$\tau\left(C\frac{Z_{k+1}-Z_k}{\tau}, \frac{Z_{k+1}-Z_k}{\tau}\right)+(AZ_{k+1}, Z_{k+1})=(AZ_k, Z_k). \quad (6.18)$$

To prove the sufficiency of condition (6.14), we must use the basic energy relation to prove the convergence of the sequence $\{\| Z_k \|\}$ to zero.

It follows from the basic energy relation and from the positive definiteness of the matrix $C=2B-\tau A$ that $(AZ_{k+1}, Z_k)\leqslant(AZ_k, Z_k)$, i.e. it follows that the sequence $\{(AZ_k, Z_k)\}$ is nonincreasing. In addition, it follows from the condition $A>0$ that this sequence is bounded below by a zero and is, therefore, convergent.

But then it follows from the basic energy relation that

$$\lim_{k\to\infty}\left(C\frac{Z_{k+1}-Z_k}{\tau}, \frac{Z_{k+1}-Z_k}{\tau}\right)=0. \quad (6.19)$$

Recall that for the positive definite matrix C there is always a $\delta>0$ such that $(CX, X)\geqslant\delta(X, X)$ for any vector X or, what is the same, $\| X \|^2\leqslant\frac{1}{\delta}(CX, X)$. The last inequality enables us to infer that from the equality of limit (6.19) to zero it follows that

$$\lim_{k\to\infty} \| Z_{k+1}-Z_k \|=0. \quad (6.20)$$

To complete the proof of the sufficiency, we must use the relation

$$B\frac{Z_{k+1}-Z_k}{\tau}+AZ_k=0$$

from which, by virtue of the existence of a limited inverse matrix A^{-1} of the positive definite matrix A, it follows that

$$Z_k=-A^{-1}\cdot\frac{B}{\tau}(Z_{k+1}-Z_k).$$

The last equation and relation (6.20) make it legitimate to conclude that $\lim\limits_{k\to\infty} \| Z_k \|=0$. We have proved the sufficiency.

To prove the necessity of conditions (6.14) under an additional assumption that the matrix B is symmetric, we use the following lemma.

Lemma. *Let C be some symmetric matrix and B, a symmetric positive definite matrix. Then the matrix C is positive definite if and only if all the eigenvalues of the problem $CX=\lambda BX$ are positive.*

To prove the lemma, we note that since the matrix B is symmetric and positive definite, there is a self-adjoint positive definite operator $B^{1/2}$, by virtue of Theorem 5.24 from 5.5.6, such that the relation $B^{1/2}\cdot B^{1/2}=B$ holds for the corresponding matrix $B^{1/2}$. Since the matrix $B^{1/2}$ is symmetric and positive definite, there is an inverse bounded and symmetric matrix which we denote by $B^{-1/2}$.

Next we note that by means of the substitution $X=B^{-1/2}\cdot Y$ and the premultiplication by the matrix $B^{-1/2}$, the problem on the eigenvalues $CX=\lambda BX$ passes into an *equivalent* problem on the eigenvalues $B^{-1/2}\cdot C\cdot B^{-1/2}\ Y=\lambda Y$, so that to prove the lemma, it is sufficient to ascertain that the obviously symmetric matrix $B^{-1/2}\cdot C\cdot B^{-1/2}$ is positive definite if and only if the matrix C is positive definite. This latter follows immediately from the fact that for any two nonzero vectors X and Y, related as $Y=B^{-1/2}\cdot X$, the equality

$$(B^{-1/2}\cdot C\cdot B^{-1/2}X,\ X)=(C\cdot B^{-1/2}X,\ B^{-1/2}X)=(CY,\ Y)$$

holds true. We have proved the lemma.

We can now pass to the proof of the necessity of conditions (6.14) of Theorem 6.2 under an additional assumption that the matrix B is symmetric.

(2) Necessity. We proceed from the following assertion from the lemma proved above: *if the matrix B is symmetric and positive definite and the matrix C is symmetric and not positive definite, then the problem on the eigenvalues $CX=\lambda BX$ has at least one nonpositive eigenvalue λ_s.*

Assume that the first of the conditions (6.14) is not satisfied, i.e. the requirement $2B-\tau A>0$ is not met.

Setting $C=2B-\tau A$ in the arguments presented above, we find that the problem on eigenvalues $(2B-\tau A)X=\lambda BX$ has at least one nonpositive eigenvalue λ_s. We designate as $X^{(s)}$ the eigenvector corresponding to λ_s and choose a zero approximation x_0 such that the condition $Z_0=X^{(s)}$ is fulfilled.

Then, rewriting the equation for the error (6.15) as $BZ_{k+1}=-BZ_k+(2B-\tau A)Z_k$, we get, setting successively, $k=0,\ 1,\ \ldots,$

$$Z_1=(-1+\lambda_s)X^{(s)},\ Z_2=(-1+\lambda_s)^2X^{(s)},\ \ldots,\ Z_k=(-1+\lambda_s)^kX^{(s)},\ldots$$

Since $-1+\lambda_s\leqslant -1$, it evidently follows that $\|Z_k\|$ does not tend to zero as $k\to\infty$.

The case when the second of conditions (6.14) is not satisfied, i.e. the condition $\tau A>0$ is not fulfilled, can be treated by analogy. In that case, in the arguments presented above we must set $C=\tau A$. We shall find that the problem $\tau AX=\lambda BX$ has at least one nonpositive eigenvalue λ_s with the eigenvector $X^{(s)}$. Choosing a zero approximation X_0 such that the equality $Z_0=X^{(s)}$ holds true and re-

writing (6.15) in the equivalent form $BZ_{k+1}=BZ_k-\tau AZ_k$, we find that

$$Z_1=(1-\lambda_s)X^{(s)},\ Z_2=(1-\lambda_s)^2X^{(s)},\ \ldots,\ Z_k=(1-\lambda_s)^kX^{(s)},\ \ldots.$$

Since $\lambda_s\leqslant 0$, it is evident that $\| Z_k \|$ does not tend to zero as $k\to\infty$. And this completes the proof of Theorem 6.2.

Let us now estimate the speed of convergence of the general implicit method of simple iteration. Proceeding in the same way as Samarsky*, we decide on the choice of the value of the parameter τ which would ensure the fastest convergence.

Assume that the matrix B is symmetric and positive definite. With the aid of such a matrix it is natural to introduce the so-called **"energy" scalar product** of two arbitrary vectors X and Y setting it equal to $(BX,\ Y)=(X,\ BY)$. We shall symbolize such a scalar product as $(X,\ Y)_B$.

With the aid of the matrix $B^{1/2}$ we can write this scalar product as $(X,\ Y)_B=(B^{1/2}B^{1/2}X,\ Y)=(B^{1/2}X,\ B^{1/2}Y)$. Using the last equation, we can easily verify the validity of the four axioms of a scalar product (see 4.1.1) for the scalar product we have introduced.

Next it is natural to introduce the **energy norm** of the vector X setting it equal to $\sqrt{(X,\ X)_B}=\sqrt{(BX,\ X)}$. We designate this energy norm as $\| X \|_B$.

Two distinct norms $\| X \|_{\mathrm{I}}$ and $\| X \|_{\mathrm{II}}$ of the same collection of vectors are said to be **equivalent** if there are positive constants γ_1 and γ_2 such that the inequalities

$$\gamma_1\| X \|_{\mathrm{I}}\leqslant\| X \|_{\mathrm{II}}\leqslant\gamma_2\| X \|_{\mathrm{I}}$$

hold true.

It should be noted that the *energy norm of the vector X and its ordinary norm are equivalent.* Indeed, the validity of the inequality $\gamma_1\| X \|\leqslant\| X \|_B$, i.e. the inequality $\gamma_1^2(X,\ X)\leqslant(BX,\ X)$ follows from positive definiteness of the matrix B, and the validity of the inequality $\| X \|_B\leqslant\gamma_2\| X \|$, i.e. the inequality $(BX,\ X)\leqslant\gamma_2^2\ \| X \|^2$ follows from the Cauchy-Bunyakovsky inequality and from estimate (6.7) (it is sufficient to set $\gamma_2^2=\| B \|$).

The equivalence of the ordinary and energy norms we have established makes it possible to assert that the *sequence* $\| X_k \|$ *converges to zero if and only if the sequence* $\| X_k \|_B$ *converges to zero.*

---

* A. A. Samarsky, *Introduction to the Theory of Difference Schemas.* Nauka, Moscow, 1971 (in Russian).

A. A. Samarsky and A. B. Gulin, *Stability of Difference Schemas,* Nauka, Moscow, 1973 (in Russian).

For further reasoning, the energy norm is more convenient than the ordinary norm.

Let us prove the following fundamental theorem.

Theorem 6.3 (Samarsky's theorem). *Suppose the matrices A and B are symmetric and positive definite, and Z_k denotes the error of the general implicit method of simple iteration. Then, for the inequality $\| Z_k \|_B \leqslant \rho^k \| Z_0 \|_B$ to hold for $\rho < 1$, it is sufficient that the condition*

$$\frac{1-\rho}{\tau} B \leqslant A \leqslant \frac{1+\rho}{\tau} B \tag{6.21}$$

holds true.

Remark. Samarsky proved that condition (6.21) was not only sufficient but also necessary for the validity of the inequality $\| Z_k \|_B \leqslant \rho^k \| Z_0 \|_B$, but we shall not dwell on this fact here.

Proof of Theorem 6.3. For the sake of convenience, we divide the proof into two steps.

1°. We shall first prove that if the symmetric and positive definite matrices A and B satisfy Samarsky's conditions (6.14), then $(BZ_{k+1}, Z_{k+1}) \leqslant (BZ_k, Z_k)$.

Multiplying scalarly equation (6.15) by $2\tau Z_{k+1} = \tau(Z_{k+1} + Z_k) + \tau(Z_{k+1} - Z_k)$, we obtain

$$(B(Z_{k+1} - Z_k), Z_{k+1} + Z_k) + (B(Z_{k+1} - Z_k), Z_{k+1} - Z_k) + \tau(AZ_{k+1}, Z_{k+1} + Z_k) + \tau(AZ_k, Z_{k+1} - Z_k) = 0.$$

In the last equation we replace AZ_k by the difference

$$\frac{1}{2} A(Z_{k+1} + Z_k) - \frac{1}{2} A(Z_{k+1} - Z_k).$$

Then, taking into account the equality $(A(Z_{k+1} - Z_k), Z_{k+1} + Z_k) = (Z_{k+1} - Z_k, A(Z_{k+1} + Z_k))$ following from the symmetry of the matrix A, we get an identity

$$(B(Z_{k+1} - Z_k), Z_{k+1} + Z_k) + ((B - \tfrac{\tau}{2} A)(Z_{k+1} - Z_k), Z_{k+1} + Z_k) + \frac{1}{2}(\tau A(Z_{k+1} + Z_k), Z_{k+1} + Z_k) = 0.$$

Taking into consideration that (by virtue of Samarsky's condition (6.14)) the operators $\tau \boldsymbol{A}$ and $\boldsymbol{B} - \frac{\tau}{2} \boldsymbol{A}$ are positive definite, we get from the last identity, an inequality

$$(B(Z_{k+1} - Z_k), Z_{k+1} + Z_k) \leqslant 0.$$

This inequality is equivalent to the inequality $(BZ_{k+1}, Z_{k+1}) \leqslant (BZ_k, Z_k)$ proved (by virtue of the identity $(BZ_{k+1}, Z_k) = (Z_{k+1}, BZ_k)$) following from the symmetry of the operator $\boldsymbol{B}$).

2°. Suppose now that Samarsky's conditions (6.21) hold true for $\rho < 1$. We shall prove the validity of the inequality $\| Z_k \|_B \leqslant \rho^k \| Z_0 \|_B$.

We set $Z_k=\rho^k V_k$. Then, evidently,

$$Z_{k+1}-Z_k=\rho^{k+1}V_{k+1}-\rho^k V_k=\rho^{k+1}(V_{k+1}-V_k)-(1-\rho)\rho^k V_k.$$

Substituting these values of Z_k and $Z_{k+1}-Z_k$ into (6.15) and cancelling by ρ^k, we get the following relation for V_k:

$$\tilde{B}\frac{V_{k+1}-V_k}{\tau}+\tilde{A}V_k=0, \tag{6.22}$$

in which $\tilde{B}=\rho B$, $\tilde{A}=A-\frac{1-\rho}{\tau}B$.

By virtue of conditions (6.21), the operators $\tilde{B}$ and $\tilde{A}$ satisfy the conditions $\tau\tilde{A}>0$, $2\tilde{B}>\tau\tilde{A}$. From these conditions and from the fact that equation (6.22) for V_k is equivalent to (6.15) for Z_k, the first step for V_k yields the following estimate: $(\tilde{B}V_{k+1}, V_{k+1})\leqslant(\tilde{B}V_k, V_k)$. Taking into account that $\tilde{B}=\rho B$, we get, from this estimate, an inequality $(BV_{k+1}, V_{k+1})\leqslant(BV_k, V_k)$.

A successive application of the indicated inequality for the numbers $k=0, 1, \ldots$ leads us to the relation $(BV_k, V_k)\leqslant(BV_0, V_0)$, and the multiplication of the last relation by ρ^{2k} leads to the final estimate* $(BZ_k, Z_k)\leqslant\rho^{2k}(BZ_0, Z_0)$.

We have thus proved the inequality $\| Z_k \|_B\leqslant\rho^k\| Z_0 \|_B$. And this completes the proof of Theorem 6.3.

In conclusion, we shall apply Samarsky's theorem 6.3 to decide on the choice of the value of the parameter τ for which the speed of the convergence is maximum. From the estimate $\| Z_k \|_B\leqslant\rho^k \|Z_0 \|_B$ which was proved in Theorem 6.3 it follows that this problem reduces to finding the value of τ for wwhich the minimum value of the function $\rho=\rho(\tau)$ is attained.

Since both matrices A and B are symmetric and positive definite, there are positive constants γ_1 and γ_2 such that the inequalities $\gamma_1 B\leqslant\leqslant A\leqslant\gamma_2 B$ hold true. We assume that the constants γ_1 and γ_2 in these inequalities are given**. Comparing the inequalities we have just written with conditions (6.21), we find that the minimum value of ρ is attained under the condition $(1-\rho)/\tau=\gamma_1$, $(1+\rho)/\tau=\gamma_2$, whence we get the optimal value $\tau=2/(\gamma_1+\gamma_2)$ and the minimum value of ρ which is equal to $(\gamma_2-\gamma_1)/(\gamma_2+\gamma_1)$.

The explicit method of simple iteration studied in 6.1.1 is a special case of the consideration presented here. All the results we have obtained are valid for that method.

In the following three subsections, we shall consider, with the aid of the implicit method of simple iteration and Samarsky's theorem

---

* We keep in mind that $z_k=\rho^k V_k$, $Z_0=V_0$.

** It is natural to call the constants γ_1 and γ_2 the **equivalence constants** of the matrices A and B. For the commuting matrices A and B the constants γ_1 and γ_2 are equal, respectively, to the least and the greatest eigenvalue of the problem $AX=\lambda BX$.

6.2, some most applicable iteration methods and establish the conditions for the convergence.

6.1.3. A modified method of simple iteration. This method can be obtained from the general implicit method of simple iteration in the special case when the stationary parameter τ is equal to unity and the matrix B is a diagonal matrix D consisting of the elements of the matrix A lying on the principal diagonal, i.e. $B=D$, where

$$D=\begin{Vmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \cdot & \cdot & & \\ 0 & 0 & \dots & a_{nn} \end{Vmatrix}. \tag{6.23}$$

In this case, we assume, of course, that the matrix A is symmetric and that all its diagonal elements $a_{11}, a_{22}, \dots, a_{nn}$ are positive (the last requirement is necessary and sufficient for the diagonal matrix $B=D$ to be positive definite).

It immediately follows from Theorem 6.2 that for the convergence of the modified method of simple iteration, for any choice of a zero approximation, it is sufficient that the conditions $2D > A$ and $A > 0$ be satisfied.

Theorem 6.1 makes it possible to express the sufficient condition for the convergence of the modified method of simple iteration in another form:

$$\| I-D^{-1}A \| < 1^* \tag{6.24}$$

(as before, the norm of the matrix is understood to be an operator form).

Since $\| I-D^{-1}A \| = \| D^{-1}(D-A) \| = \| D^{-1}(A-D) \|$, the sufficient condition for convergence (6.24) can be rewritten in the equivalent form

$$\| D^{-1}(A-D) \| < 1. \tag{6.25}$$

Inequality (6.25) permits us to obtain various sufficient conditions for the convergence of the modified method of simple iteration.

Note, first of all, that if alongside the operator norm of matrix (6.2.) (which we denote by $\| A \|$ as before), we introduce the so-called spherical norm of that matrix which is designated as $\| A \|_{\text{sph}}$ and defined by the equation $\| A \|_{\text{sph}} = \left[\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2 \right]^{1/2}$, then, as

* We take into account that in the case being considered we must take the matrix $\tilde{A}$ defined by the formula $\tilde{A}=B^{-1}A$ instead of the matrix A and set $B=D$, $\tau=1$.

was proved in 4.4 (see formula (4.28)), the inequality*

$$\| AX \| \leqslant \| A \|_{\text{sph}} \| X \| \tag{6.26}$$

is valid for any vector X of the space E^n.

It immediately follows from (6.26) and (6.5) that the operator and the spherical norm of the matrix are related as $\| A \| \leqslant \| A \|_{\text{sph}}$.

Thus, by virtue of (6.25), the sufficient condition for the convergence of the modified method of simple iteration can be expressed by the inequality $\| D^{-1}(A-d) \|_{\text{sph}} < 1$ which can be written out as $\sum\limits_{i=1}^{n} \sum\limits_{\substack{j=1 \\ j \neq 1}}^{n} \frac{a_{ij}^2}{a_{ii}^2} < 1$.

It should next be pointed out that when defining the operator norm (6.5) of the matrix A we proceeded from the ordinary (so-called **spherical**) norm of the vector $X = \begin{Vmatrix} x_1 \\ \cdots \\ x_n \end{Vmatrix}$ which is equal to $\| X \| = \left[\sum\limits_{i=1}^{n} x_i^2 \right]^{1/2}$.

Another two norms of the vector X are often introduced: the so-called **cubic norm** defined by the equation $\| X \|_{\text{cub}} = \max\limits_{1 \leqslant i \leqslant n} | x_i |$ and the so-called **octahedral norm** defined by the equation $\| X \|_{\text{oct}} = \sum\limits_{i=1}^{n} | x_i |$. If the norm of the vector in definition (6.5) of the operator norm of the matrix A is understood as its cubic or octahedral norm respectively, then relation (6.5) leads to the definition of the **cubic** and the **octahedral operator norm** of the matrix A respectively.

We can prove that the cubic and the octahedral operator norm of matrix (6.2) can be expressed in terms of the elements of that matrix as follows:

$$\| A \|_{\text{cub}} = \max_{1 \leqslant i \leqslant n} \sum_{j=1}^{n} | a_{ij} |,$$

$$\| A \|_{\text{oct}} = \max_{1 \leqslant j \leqslant n} \sum_{i=1}^{n} | a_{ij} |.$$

---

* In this inequality the norm of the vector $X = \begin{Vmatrix} x_1 \\ \cdots \\ x_n \end{Vmatrix}$ is understood as so-called spherical norm $\| X \| = \left[\sum\limits_{i=1}^{n} x_i^2 \right]^{1/2}$.

The verbatum repetition of the reasoning for the replacement of spherical norms by cubic and octahedral norms respectively leads to the sufficient condition for the convergence of the modified method of simple iteration expressed by relation (6.25) in which the norm of the matrix should be understood as its cubic or octahedral operator norm respectively.

This leads to the following two conditions each of which is a sufficient condition for the convergence of the modified method of simple iteration:

$$\sum_{\substack{j=1\\ j\neq i}}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| < 1 \quad (\text{for } i=1,\ 2,\ \ldots,\ n),$$

$$\sum_{\substack{i=1\\ i\neq 1}}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| < 1 \quad (\text{for } j=1,\ 2,\ \ldots,\ n).$$

6.1.4. Seidel's method. Let us represent the symmetric matrix (6.2) as the sum of three matrices $A=D+L+U$, where D is the diagonal matrix (6.23) and L and U are, respectively, the strictly left and strictly right matrices having the form

$$L=\begin{Vmatrix} 0 & 0 & & 0 \\ a_{21} & 0 & & 0 \\ \cdot & \cdot & & \cdot \\ a_{n1} & a_{n2} & \ldots & 0 \end{Vmatrix}, \quad U=\begin{Vmatrix} 0 & a_{12} & \ldots & a_{1n} \\ 0 & 0 & & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \ldots & 0 \end{Vmatrix}$$

and satisfying the condition $L'=U$.

Seidel's method can be obtained from the general implicit method of simple iteration in the special case when the stationary parameter τ is equal to unity and the matrix B is equal to the sum $D+L$. Thus the successive iterations in Seidel's method are defined by the relation

$$(D+L)(X_{k+1}-X_k)+AX_k=F.$$

Let us prove that *Seidel's method converges for any symmetric and positive definite matrix A.*

By Theorem 6.2, it is sufficient to prove that the condition

$$2(D+L)>A \tag{6.27}$$

holds for any such matrix A.

To prove (6.27), we pay attention to the fact that

$$\begin{aligned}(2(D+L)X,\ X)&=(DX,\ X)+(DX,\ X)+(LX,\ X)+(LX,\ X)\\ &=(DX,\ X)+(DX,\ X)+(LX,\ X)+(X,\ UX)\\ &=(DX,\ X)+(AX,\ X)\end{aligned}$$

for any vector X.

Thus, to prove inequality (6.27), it is sufficient to verify that the matrix D is positive definite but this follows immediately from the fact that all the elements of the symmetric and positive definite matrix A which lie on the principal diagonal are positive*. We have thus proved the convergence of Seidel's method.

6.1.5. The method of overrelaxation. This method can be derived from the general method of simple iteration in the special case when $\tau=\omega$, $B=D+\omega L$ and the parameter ω is chosen such that the eigenvalue of the matrix $I-\omega(D+\omega L)^{-1}A$, carrying out the passage from the kth iteration to the $(k+1)$th, with the largest absolute value, is the least.

Let us prove that *if the matrix A is symmetric and positive definite, then for the overrelaxation method to converge, it is sufficient that the condition* $0<\omega<2$ *be satisfied.*

By Theorem 6.2, for the method to converge, it is sufficient that the conditions $\omega>0$, $2(D+\omega L)>\omega A$ be satisfied.

For any vector X, the second condition leads to the inequality

$$(2(D+\omega L)X,\ X)>(\omega AX,\ X). \tag{6.28}$$

The last inequality is equivalent to each of the inequalities in the following chain:

$$(2DX,\ X)+(\omega LX,\ X)+(\omega LX,\ X)>(\omega AX,\ X),$$

$$(2-\omega)(DX,\ X)+(\omega DX,\ X)+(\omega LX,\ X)+(X,\ \omega UX)>(\omega AX,\ X),$$

$$(2-\omega)(DX,\ X)>0.$$

We infer from the last inequality and from the positive definiteness of D that (6.28) is valid for $2-\omega>0$, i.e. for $\omega<2$. We have thus proved that the conditions $0<\omega<2$ ensure the convergence of the method of overrelaxation.

6.1.6. The case of a nonsymmetric matrix *A*. When the matrix A is nonsymmetric, we can premultiply matrix equation (6.1) by the matrix A' and replace equation (6.1) by the equation $\tilde{A}X=\tilde{F}$ in which $\tilde{F}=A'F$, $\tilde{A}=A'A$ so that the matrix $\tilde{A}$ will be symmetric and (as can be easily verified) positive definite.

6.1.7. Chebyshev's iterative method.** Everywhere above, when we considered the general implicit method of simple iteration, we supposed that the iteration parameter τ assumed one and the same constant value. An idea naturally arises to consider a more general case when the values of the iteration parameter in the indicated method depend on the number k of iteration. In that case, the sequence of the

---

* It is sufficient to note that if the kth coordinate of the vector X is equal to unity and all the other coordinates are zero, then $(Ax,\ X)=a_{kk}>0$.

** Chebyshev, Pafnuty Lvovich (1821-1894), an eminent Russian mathematician.

iterations is defined not by the relation

$$B\frac{X_{k+1}-X_k}{\tau}+AX_k=F, \tag{6.13}$$

but by a more general relation

$$B\frac{X_{k+1}-X_k}{\tau_{k+1}}+AX_k=F.$$

As before, B is a certain easily convertible square matrix of order n. With such a choice of the iteration sequence, we get the following relation for the error $Z_k=X_k-X$ of the iteration scheme:

$$B\frac{Z_{k+1}-Z_k}{\tau_{k+1}}+AZ_k=0. \tag{6.15*}$$

Let us assume that both matrices A and B are symmetric and positive definite. Then, as was noted above, there are positive constants γ_1 and γ_2 such that $\gamma_1 B\leqslant A\leqslant\gamma_2 B$. We assume that these constants are specified and recall once again that they are equal to the least and the greatest eigenvalue of the problem $AX=\lambda BX$ respectively. Let us evaluate the energy norm of the error $\| Z_k \|_B$.

Recall once again that for the symmetric and positive definite matrix B there is a symmetric and positive definite matrix $B^{1/2}$ such that $B^{1/2}\cdot B^{1/2}=B$. As before, we agree to use the symbol $B^{-1/2}$ for the inverse of the matrix $B^{1/2}$.

To estimate the norm of the error Z_k, we make a replacement setting $Z_k=B^{-1/2}\cdot V_k$. With such a replacement, the relation for the error Z_k passes into the following relation for V_k:

$$V_{k+1}=(I-\tau_{k+1}C)\cdot V_k \quad (k=0,\ 1,\ 2,\ \ldots),$$

where C denotes a matrix of the form $C=B^{-1/2}\cdot A\cdot B^{-1/2}$. Let us verify that the square of the ordinary norm of the vector V_k is equal to the square of the energy norm of the vector z_k. Indeed,

$$\| V_k \|^2=(V_k,\ V_k)=(B^{1/2}Z,\ B^{1/2}Z)=(BZ_k,\ Z_k)=\| Z_k \|_B^2.$$

Thus, to estimate the energy norm Z_k, it is sufficient to estimate the ordinary norm V_k.

Let us estimate the norm $\| V_k \|$. We note first of all that by means of the replacement $X=B^{-1/2}\cdot Y$, we can get inequalities $\gamma_1(Y,\ Y)\leqslant(CY,\ Y)\leqslant\gamma_2(Y,\ Y)$ from the inequalities $\gamma_1(BX,\ X)\leqslant(AX,\ X)\leqslant\gamma_2(BX,\ X)$. The inequalities $\gamma_1(Y,\ Y)\leqslant(CY,\ Y)\leqslant\gamma_2(Y,\ Y)$ are equivalent to $\gamma_1 I\leqslant C\leqslant\gamma_2 I$. Since, in addition, the matrix $C=B^{-1/2}\cdot A\cdot B^{-1/2}$ is symmetric, all the eigenvalues of that matrix are real and lie on the interval $[\gamma_1,\ \gamma_2]$. Writing consecutively the relation

$V_{k+1}=(I-\tau_{k+1}\cdot C)V_k$ for the numbers $k=0, 1, \ldots,$ we arrive at an equation

$$V_k=\prod_{j=1}^{k}(I-\tau_j C)\cdot V_0$$

which immediately yields

$$\| V_k \| \leqslant \left\| \prod_{j=1}^{k}(I-\tau_j C)\right\| \cdot \| V_0 \|.$$

But then it follows from equation $\| V_k \| = \| Z_k \|_B$ that $\| Z_k \|_B \leqslant q^k \| Z_0 \|_B$, where $q_k = \left\| \prod_{j=1}^{k}(I-\tau_j C) \right\|$. Consequently, the iterative process converges provided that the sequence $\{q_k\}$ tends to zero, and the convergence is the faster the smaller the quantities q_k.

Since every value of q_k is a function of the parameters τ_1, $\tau_2, \ldots, \tau_k$, a problem arises of constructing an optimal collection of iteration parameters from the condition of minimum of q_k for a fixed k. Let us pass to the solution of this problem.

We assume that all the eigenvalues λ_s of the matrix C lie on the given interval $[\gamma_1, \gamma_2]$. Taking the symmetry of the matrix C into account, we arrive at the following optimization problem: it is required to find

$$\min_{\{\tau_j\}} q_k(\tau_1, \tau_2, \ldots, \tau_k)=\min_{\{\tau_j\}}\left\| \prod_{j=1}^{k}(I-\tau_j C)\right\| = \min_{\{\tau_j\}}\left\{ \max_s \left| \prod_{j=1}^{k}(I-\tau_j\lambda_s)\right| \right\}.$$

Since all λ_s lie on the interval $[\gamma_1, \gamma_2]$, by extending the domain over which the maximum is taken, we find that

$$\min_{\{\tau_j\}} q_k(\tau_1, \tau_2, \ldots, \tau_k)\leqslant \min_{\{\tau_j\}}\left\{ \max_{\gamma_1\leqslant t\leqslant\gamma_2} \left| \prod_{j=1}^{k}(I-\tau_j t)\right| \right\}.$$

The rough problem we have obtained has a simpler solution. Besides, when solving such a problem, we do not use the information on the actual position of the eigenvalues λ_s on the interval $[\gamma_1, \gamma_2]$ but only take into consideration the boundaries of that interval. Such an approach makes it possible to construct a collection of optimal parameters for matrices of arbitrary structure.

Let us solve the rough problem of optimization. We set $P(t)=\prod_{j=1}^{k}(1-\tau_j t)$ and note that the polynomial $P(t)$ satisfies the condition of valuation $P(0)=1$. By changing the variable

$$t=\frac{1}{2}[\gamma_2+\gamma_1-S(\gamma_2-\gamma_1)]=\frac{\gamma_2+\gamma_1}{2}\left(1-S\frac{\gamma_2-\gamma_1}{\gamma_2+\gamma_1}\right)=\frac{1-S\rho_0}{\tau_0},$$

where $\rho_0=\frac{\gamma_2-\gamma_1}{\gamma_2+\gamma_1}$, $\tau_0=\frac{2}{\gamma_1+\gamma_2}$, we map the interval $\gamma_1\leqslant t\leqslant\gamma_2$ into the interval $-1\leqslant S\leqslant 1$, the point $t=0$ passing into the point $S=S_0=\frac{1}{\rho_0}>1$.

With such a change, the optimization problem in question passes into the following problem: *among all the polynomials $\bar{P}_k(S)$ of degree k which satisfy the valuation condition $\bar{P}_k\left(\frac{1}{\rho_0}\right)=1$, find the polynomial for which* $\max\limits_{|S|\leqslant 1}|\bar{P}(S)|$ *is minimal.*

As we know, this is Chebyshev's polynomial $P_k(S)=T_k(S)\cdot[T_k(S_0)]^{-1}$, where

$$T_k(S)=\begin{cases}\cos(k\arccos S) & \text{for } |S|\leqslant 1,\\ \frac{1}{2}[(S+\sqrt{S^2-1})^k+(S-\sqrt{S^2-1})^k] & \text{for } |S|>1.\end{cases}$$

Since $\max\limits_{|S|\leqslant 1}|T_k(S)|=1$, we have

$$\min_{\{\tau_i\}}\ \max_{\gamma_1\leqslant t\leqslant\gamma_2}|P_k(t)|=\frac{1}{T_k(S_0)},$$

with $\frac{1}{T_k(S_0)}=q_k=\frac{2\rho_1^k}{1+\rho_1^{2k}}$, where $\rho_1=\frac{\sqrt{\gamma_2}-\sqrt{\gamma_1}}{\sqrt{\gamma_2}+\sqrt{\gamma_1}}$.

To calculate the optimal collection of parameters, we proceed from the equation

$$P_k(t)=\prod_{j=1}^{k}(1-\varepsilon_j t)=q_kT_k\left(\frac{1-\varepsilon_0 t}{\rho_0}\right)$$

$\left(\text{we have taken into account that } S=\frac{1-\tau_0 t}{\rho_0}\right)$. We equate the roots of the polynomials appearing on the left-hand and right-hand sides of this equation. Since the polynomial $P_k(t)$ has the roots $t_j=1/\tau_j$ $(j=1, 2; \ldots, k)$ and the polynomial $T_k(S)$ has the roots $S_j=\cos\frac{2j-1}{2k}\pi$ $(j=1,2,\ldots,k)$, we take into account that $t=\frac{1-\rho_0 S}{\tau_0}$ and get $\frac{1}{\tau_j}=\frac{1-S_j\rho_0}{\tau_0}$ $(j=1, 2, \ldots, k;$ S_j were defined above).

Thus, the values $\tau_j=\frac{\tau_0}{1+S_0S_j}$, where $S_j=\cos\frac{2j-1}{2k}\pi$, $j=1, 2,\ldots, k$ are the optimal values of the iteration parameters.

The iteration process with the indicated collection of parameters is known as **Chebyshev's process.**

We arrive at the following theorem.

Theorem 6.4. *If the matrices A and B are symmetric and positive definite and if $\gamma_1 B \leqslant A \leqslant \gamma_2 B$, then Chebyshev's iterative process converges, and the estimate*

$$\| Z_k \|_B \leqslant q_k \| Z_0 \|_B,$$

where $q_k = \frac{2\rho_1^k}{1+\rho_1^{2k}}$ *for* $\rho_1 = \frac{\sqrt{\gamma_2}-\sqrt{\gamma_1}}{\sqrt{\gamma_2}+\sqrt{\gamma_1}}$ *is valid for the error Z_k after k iterations.*

If we take the requirement $\| Z_k \|_B < \varepsilon \| Z_0 \|_B$ as the condition for ending the process for the preassigned ε-accuracy, then Theorem 6.4 yields the following estimate for k iterations: $k \geqslant k_0(\varepsilon) = \frac{\ln(\varepsilon/2)}{\ln \rho_1}$. Comparing this estimate with the obtained estimate $\bar{k} \geqslant \bar{k}_0(\varepsilon) = \frac{\ln \varepsilon}{\ln \rho_0}$ of the number of iterations for the method of simple iteration, we find, provided that the quantity $\xi = \gamma_2/\gamma_1$ is small, that $k_0(\varepsilon) \approx \frac{\ln(2/\varepsilon)}{2\sqrt{\xi}}$, $\bar{k}_0(\varepsilon) \approx \frac{\ln(1/\varepsilon)}{2\xi}$. The comparison of these two estimates shows that Chebyshev's method has certain advantages (when the quantity $\xi = \gamma_1/\gamma_2$ is small).

Chebyshev's method we have described have been known since the beginning of 1950s. It is sometimes called **Richardson's method.**

It should be pointed out that we have studied the method for an ideal computing process with an infinite number of digits whereas calculations on computers are carried out with a finite number of digits and, therefore, there are numbers which are computer infinity M_∞ and computer zero. If, in the process of computer calculations, there appears a number M exceeding M_∞ an emergency shut-down of the computer occurs.

From the point of view of an ideal computing process, the values of the iteration parameters τ_j can be ordered in any way (by any of the $k!$ methods). From the viewpoint of an ideal computing process, any two sequences of iteration parameters $\{\tau_j\}$ are equivalent since the ε-accuracy they require can be attained in the same number of iterations.

But when calculations are carried out with the aid of computers, various sequences of parameters $\{\tau_j\}$ are not equivalent. For some sequences of values $\{\tau_j\}$ an emergency shut-down of the computer may occur because of a growth of intermediate values. For other sequences of values $\{\tau_j\}$ no emergency shut-down occurs, but because of the nonmonotonic tendency of the error Z_k to zero, i.e. owing to the fact that the norm of the matrix $I - \tau_j C$ of transition from the $(j-1)$ th

iteration to the jth iteration can be greater than unity, the estimate established for an ideal situation is not valid for this error.

Because of the indicated circumstances, a theoretical problem arises which consists in indicating the best law of ordering the values $\{\tau_j\}$ for which the influence of the errors of rounding-off is the least for Chebyshev's method.

You can find the exhaustive solution of this problem in Samarsky's book *The Theory of Difference Schema,* Nauka, Moscow, 1977 (in Russian).

6.2. Solving a Complete Problem of Eigenvalues by the Rotation Method

For the sake of simplicity, we shall first consider a real symmetric matrix A defined by equation (6.2). Note that the seeking of all the eigenvalues and eigenvectors of this matrix reduces to seeking an orthogonal matrix T such that the product

$$D = T'AT \tag{6.29}$$

is a diagonal matrix. Indeed, if such an orthogonal matrix T is found, the diagonal elements of the matrix D are the eigenvalues of the matrix A and the columns of the matrix T are the respective eigenvectors of the matrix A*.

Let us introduce into consideration the spherical norm of the matrix A:

$$\| A \|_{\text{sph}} + \left[\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2 \right]^{1/2}.$$

Then, evidently, the inequality

$$\sum_{i=1}^{n} a_{ii}^2 \leqslant \| A \|_{\text{sph}}^2 \tag{6.30}$$

is valid for the diagonal elements of the matrix A, and this inequality passes into an exact equality only in the case when the matrix A is diagonal.

It should be pointed out that *under the orthogonal transformation of the matrix A* (i.e. under the transformation of the form $\tilde{A} = UAR$, where U and R are orthogonal matrices), *the spherical norm*

---

* To prove this fact we denote the diagonal elements of the matrix D *by* $\lambda_1, \lambda_2, \ldots, \lambda_n$ and set $e_k = \| e_k^i \|$, where the elements e_k^i of the column e_k satisfy the condition $e_k^i = 0$ for $k \neq i$ and $e_k^k = 1$. Then, evidently. $De_k = \lambda_k e_k$, i.e. $T'ATe_k = \lambda_k e_k$ and since $T' = T^{-1}$, we have $ATe_k = \lambda_k T_k$. Consequently, Te_k are the eigenvectors of the matrix A.

*of that matrix does not change**. Hence it follows that the difference between transformation (6.29) and all the orthogonal transformations of the matrix A is that the former makes the sum of the squares of the diagonal elements of the transformed matrix maximal and the sum of the squares of all the off-diagonal elements of that matrix minimal.

The rotation method is an iterative method for which the matrix T indicated above can be found as the limit of the infinite product of elementary rotation matrices each of which has the form

$$T_{ij}(\varphi)=\begin{Vmatrix} 1 & & & & & & & & & 0 \\ & \ddots & & & & & & & & \\ & & 1 & & & & & & & \\ & & & \cos\varphi\ldots & & & & -\sin\varphi\ldots & & \\ & & & \vdots & 1 & & & \vdots & & \\ & & & \vdots & & \ddots & & \vdots & & \\ & & & & & & 1 & & & \\ & & & \sin\varphi\ldots & & & & \cos\varphi\ldots & & \\ & & & & & & & & 1 & \\ & & & & & & & & & \ddots \\ 0 & & & & & & & & & 1 \end{Vmatrix} \begin{matrix} \\ \\ \\ \text{(the } i\text{th row)} \\ \\ \\ \\ \text{(the } j\text{th row).} \\ \\ \\ \\ \end{matrix} \tag{6.31}$$

In the whole, the method of rotations consists in the construction of the sequence of matrices

$$A,\ A_1,\ A_2,\ \ldots,\ A_\nu,\ A_{\nu+1},\ \ldots, \tag{6.32}$$

each successive matrix being obtained from the preceding matrix by means of the elementary step of the form $A_{\nu+1}=T'_{ij}A_\nu T_{ij}$.

If we simplify the notation by omitting the index ν and consider one step $\tilde{A}=T'_{ij}AT_{ij}$ performed with the aid of matrix (6.31), then we get the following expressions for the elements $\tilde{a}_{ij}$ of the transformed matrix $\tilde{A}$ in terms of the elements a_{ij} of the matrix A

$\tilde{a}=a_{kl}$ for $k\neq i,\ j,\ l\neq i,\ j,$
$\tilde{a}_{il}=a_{il}\cos\varphi+a_{jl}\sin\varphi$ for $l\neq i,\ j,$
$\tilde{a}ji=-a_{il}\sin\varphi+a_{jl}\cos\varphi$ for $l\neq i,\ j,$
$\tilde{a}_{li}=a_{li}\cos\varphi+a_{lj}\sin\varphi$ for $l\neq i,\ j,$
$\tilde{a}_{lj}=-a_{li}\sin\varphi+a_{lj}\cos\varphi$ for $l\neq i,\ j,$
$\tilde{a}_{ii}=(a_{ii}\cos\varphi+a_{ji}\sin\varphi)\cos\varphi+(a_{ij}\cos\varphi+a_{jj}\sin\varphi)\sin\varphi.$
$\tilde{a}_{ji}=(-a_{ii}\sin\varphi+a_{ji}\cos\varphi)\cos\varphi+(-a_{ij}\sin\varphi+a_{jj}\cos\varphi)\sin\varphi,$

* Indeed, if $\tilde{A}=UAR$ and the symbol tr C denotes the sum of all the elements of the matrix C lying on its principal diagonal, then $\|\tilde{A}\|^2_{\text{sph}}=\operatorname{tr}(\tilde{A}',\ \tilde{A})=\operatorname{tr}(R'A'U'UAR)=\operatorname{tr}(R'A'AR)=\|AR\|^2_{\text{sph}}=\|(AR)'\|^2_{\text{sph}}=\|R'A'\|^2_{\text{sph}}=\operatorname{tr}(ARR'A')=\operatorname{tr}(AA')=\|A'\|^2_{\text{sph}}=\|A\|^2_{\text{sph}}$.

$$\tilde{a}_{jj} = -(-a_{ii}\sin\varphi + a_{ji}\cos\varphi)\sin\varphi + (-a_{ij}\sin\varphi + a_{jj}\cos\varphi)\cos\varphi,$$
$$\tilde{a}_{ij} = -(a_{ii}\cos\varphi + a_{ji}\sin\varphi)\sin\varphi + (a_{ij}\cos\varphi + a_{jj}\sin\varphi)\cos\varphi. \tag{6.33}$$

Relations (6.33) and the condition of symmetry of the matrix A yield the following equality that can be easily verified:

$$\sum_{\substack{k=1\\k\neq l}}^{n}\sum_{l=1}^{n}\tilde{a}_{kl}^2 = \sum_{\substack{k=1\\k\neq l}}^{n}\sum_{i=1}^{n}a_{kl}^2 - 2a_{ij}^2$$
$$+\frac{1}{2}[(a_{jj}-a_{ii})\sin 2\varphi + 2a_{ij}\cos 2\varphi]^2. \tag{6.34}$$

It follows from this equation that for the maximum reduction of the sum of the squares of all the off-diagonal elements, it is necessary to choose matrix (6.31) such that the following **two requirements** are met.

(1) it is required to choose the numbers i and j such that the square of the element a_{ij}^2 is the greatest among the squares of all the off-diagonal elements of the matrix A, i.e. to subordinate the choice of the numbers i and j to the condition $a_{ij}^2 = \max\limits_{\substack{1\leqslant k\leqslant n\\1\leqslant l\leqslant n\\k\neq l}} a_{kl}^2$.

(2) the angle of rotation φ in the matrix (6.31) should be chosen such that the equality

$$(a_{jj}-a_{ii})\sin 2\varphi + 2a_{ij}\cos 2\varphi = 0 \tag{6.35}$$

holds true.

Equation (6.35) defines uniquely the angle φ satisfying the conditions

$$\tan 2\varphi = \frac{2a_{ij}}{a_{ii}-a_{jj}}, \quad |\varphi| \leqslant \frac{\pi}{4}. \tag{6.36}$$

This equation makes it possible to calculate $\cos\varphi$ and $\sin\varphi$ by the formulas

$$\cos\varphi = \left\{\frac{1}{2}[1+(1+p^2)^{-1/2}]\right\}^{1/2},$$
$$\sin\varphi = \operatorname{sgn} p\left\{\frac{1}{2}[1-(1+p^2)^{-1/2}]\right\}^{1/2},$$

where $p = 2a_{ij}/(a_{ii}-a_{jj})$.

Note that if we have chosen matrix (6.31) such that the indicated requirements (1) and (2) are met, then equation (6.34) acquires the following form:

$$\sum_{\substack{k=1\\k\neq l}}^{n}\sum_{l=1}^{n}\tilde{a}_{kl}^2 = \sum_{\substack{k=1\\k\neq l}}^{n}\sum_{l=1}^{n}a_{kl}^2 - 2a_{ij}^2, \tag{6.37}$$

in which a_{ij} is an off-diagonal element of the matrix having the largest absolute value.

Now we can say more exactly that the rotation method consists in constructing a sequence of matrices (6.32), each successive matrix being obtained from the preceding matrix by means of the orthogonal transformation $A_{\nu+1}=T'_{ij}A_\nu T_{ij}$ in which the matrix $T_{ij}=T_{ij}(\varphi)$ is chosen such that the two requirements indicated above are met*.

Let us prove the convergence of the method of rotations. We denote by S_ν^2 the sum of the squares of all the off-diagonal elements of the matrix A_ν and by $a_{i_\nu j_\nu}^{(\nu)}$ the off-diagonal element of this matrix which is the largest in the absolute value.

Then, by virtue of (6.37), there holds an equality

$$S_{\nu+1}^2=S_\nu^2-2\,[a_{i_\nu j_\nu}^{(\nu)}]^2. \tag{6.38}$$

Furthermore, since the total number of the off-diagonal elements of the matrix A is equal to $n(n-1)$, and $a_{i_\nu j_\nu}^{(\nu)}$ is the greatest of these elements in absolute value, then the equality

$$\left[a_{i_\nu j_\nu}^{(\nu)}\right]^2 \geqslant \frac{S_\nu^2}{n(n-1)} \tag{6.39}$$

holds true. But it follows from (6.38) and (6.39) that

$$S_{\nu+1}^2 \leqslant S_\nu^2\left[1-\frac{2}{n(n-1)}\right]. \tag{6.40}$$

Using successively inequality (6.40) written for the numbers 0, 1, ..., ν and designating as $S_0^2=S_0^2(A)$ the sum of the squares of all the off-diagonal elements of the system matrix A, we find that

$$S_{\nu+1}^2 \leqslant S_0^2(A)\left[1-\frac{2}{n(n-1)}\right]^{\nu+1}. \tag{6.41}$$

It immediately follows from inequality (6.41) that $\lim\limits_{\nu\to\infty} s_{\nu+1}^2=0$, and this proves the convergence of the rotation method.

The diagonal elements of the matrix A_ν are taken as approximate values of the eigennumbers of the matrix A and the columns of the matrix $T_{i_1j_1}\cdot T_{i_2j_2}\ldots T_{i_\nu j_\nu}$ are taken as approximate eigenvectors of the matrix A.

More exact results were obtained by Voevodin**. For the case when an arbitrary (not necessarily symmetric) matrix A does not have jor-

* At each step the numbers i and j are chosen such that the off-diagonal element of the matrix A_ν with these numbers should have the largest absolute value.

** V. V. Voevodin, *Numerical Methods of Algebra. Theory and Algorithms.* Nauka, Moscow, 1966 (in Russian).

dan blocks and all its off-diagonal elements are values of order ε and are small as compared to the number $\rho = \min\limits_{\lambda_i \neq \lambda_j} |\lambda_i - \lambda_j|$, Voevodin obtained the following estimates:

(a) the estimate $\lambda_i = a_{ii} + \sum\limits_{p=1}^{n} \frac{a_{ip}a_{pi}}{a_{ii} - a_{pp}} + O(\varepsilon^3)$ for the eigenvalues

(we exclude, from the indicated sum, the values p, belonging to the set R_l, of the numbers $j = 1, 2, \ldots, n$ for which $\lambda_j = \lambda_i$);

(b) if T is a matrix whose columns are the eigenvectors of the matrix A and $T = I + H$, where I is an identity matrix, then the estimates

$$h_{ij} = \begin{cases} 0 & \text{if} \quad \lambda_i = \lambda_j \\ \dfrac{a_{ij}}{a_{jj} - a_{ii}} + O(\varepsilon^3) & \text{if} \quad \lambda_i \neq \lambda_j \end{cases}$$

are valid for the elements h_{ij} of the matrix H.

If A is a **complex Hermitian** matrix, then we must take the unitary matrix

$$T_{ij}(\varphi, \psi) = \begin{array}{c} \left\| \begin{array}{ccccccccc} 1 & & & & & & & & 0 \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & \cos\varphi & \cdots & & -\sin\varphi e^{i\psi} & \cdots & \\ & & & \vdots & 1 & & \vdots & & \\ & & & & & \ddots & & & \\ & & & & & 1 & \vdots & & \\ & & & \sin\varphi e^{-i\psi} & & & \cos\varphi & \cdots & \\ & & & & & & 1 & & \\ & & & & & & & \ddots & \\ 0 & & & & & & & & 1 \end{array} \right\| \end{array} \begin{array}{l} \\ \\ \\ \text{(the } i\text{th row)}, \\ \\ \\ \\ \text{(the } j\text{th row)}, \\ \\ \\ \\ \end{array} \tag{6.42}$$

instead of matrix (6.31). Then, instead of equation (6.34), we arrive at the equation

$$\sum_{k=1}^{n} \sum_{\substack{l=1 \\ k \neq l}}^{n} |\tilde{a}_{kl}|^2 = \sum_{k=1}^{n} \sum_{\substack{l=1 \\ k \neq l}}^{n} |a_{kl}|^2 - 2|a_{ij}|^2$$

$$+ 2|a_{ij}| \cdot |\cos^2\varphi \cdot e^{i\alpha} - \sin^2\varphi \cdot e^{i(2\psi - \alpha)}$$

$$+ (a_{jj} - a_{ii}) \cos\varphi \sin\varphi e^{i\psi}|^2$$

in which α denotes the argument of the complex number a_{ij}.

To make the sum of the squares of the absolute values of the off-diagonal elements as small as possible, we must take the numbers i and j of matrix (6.42) such that the element a_{ij} is the greatest off-diagonal element of the matrix A in absolute value and subject the choice of the angles φ and ψ to the condition

$$|a_{ij}|\,(\cos^2\varphi\cdot e^{i\alpha}-\sin^2\varphi\cdot e^{i(2\psi-\alpha)}$$
$$+(a_{jj}-a_{ii})\cos\varphi\cdot\sin\varphi\cdot e^{i\psi})=0.$$

The last condition leads to relations

$$\Psi=\arg a_{ij},\quad \tan 2\varphi=\frac{2\,|a_{ij}|}{a_{ii}-a_{jj}},\quad |\varphi|\leqslant\frac{\pi}{4}.$$

The convergence of the rotation method can be proved in the same way as for the case of a real matrix.

Chapter 7

BILINEAR AND QUADRATIC FORMS

In this chapter we study **bilinear forms** defined in a real linear space, i.e. the number functions of two vector arguments which are linear with respect to each argument. We study in detail the so-called **quadratic forms** which are bilinear forms defined for the coinciding values of their arguments. We also consider some applications of the theory of bilinear and quadratic forms.

7.1. Bilinear Forms

7.1.1. The concept of a bilinear form. The notion of a bilinear form in an arbitrary linear space was introduced in Chapter 5. However, to make the presentation of the material more convenient, we recall here some definitions and simple assertions.

Definition 1. *The number function $A(\boldsymbol{x}, \boldsymbol{y})$, whose arguments are various vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ of the real linear space L, is called a* ***bilinear form*** *if the relations*

$$\left.\begin{aligned} A(\boldsymbol{x}+\boldsymbol{z}, \boldsymbol{y}) &= A(\boldsymbol{x}, \boldsymbol{y})+A(\boldsymbol{z}, \boldsymbol{y}), \\ A(\boldsymbol{x}, \boldsymbol{y}+\boldsymbol{z}) &= A(\boldsymbol{x}, \boldsymbol{y})+A(\boldsymbol{x}, \boldsymbol{z}), \\ A(\lambda\boldsymbol{x}, \boldsymbol{y}) &= \lambda A(\boldsymbol{x}, \boldsymbol{y}), \\ A(\boldsymbol{x}, \lambda\boldsymbol{y}) &= \lambda A(\boldsymbol{x}, \boldsymbol{y}) \end{aligned}\right\} \tag{7.1}$$

hold for any vectors $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$ from L and real number λ.

In other words, *a bilinear form is a number function $A(\boldsymbol{x}, \boldsymbol{y})$ of two vector arguments $\boldsymbol{x}$ and $\boldsymbol{y}$ which is defined on various vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ of the real linear space L and is linear with respect to each argument**.

The simplest example of a bilinear form is the product of two linear forms $f(\boldsymbol{x})$ and $g(\boldsymbol{y})$ defined on the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ of the linear space L.

Definition 2. *The bilinear form $A(\boldsymbol{x}, \boldsymbol{y})$ is said to be* ***symmetric***

* It is often said in this case that the bilinear form $A(\boldsymbol{x}, \boldsymbol{y})$ *is given on the linear* space L.

(***skew-symmetric***) *if the relations*

$$A(\boldsymbol{x}, \boldsymbol{y}) = A(\boldsymbol{y}, \boldsymbol{x}) \quad (A(\boldsymbol{x}, \boldsymbol{y}) = -A(\boldsymbol{y}, \boldsymbol{x})) \tag{7.2}$$

hold for any vectors $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *of the linear space* L.

The following *statement* holds true: *any bilinear form can be represented as the sum of a symmetric and a skew-symmetric bilinear form* (see 5.9.1).

7.1.2. Representation of a bilinear form in a finite-dimensional linear space. Suppose a bilinear form $B(\boldsymbol{x}, \boldsymbol{y})$ is given in an n-dimensional linear space L. Let us decide on the representation of the form $B(\boldsymbol{x}, \boldsymbol{y})$ in the case when a definite basis $e = (\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)$ is specified in L.

The following assertion holds true.

Theorem 7.1. *For the basis* $e = (\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots \boldsymbol{e}_n)$ *specified in an* n-*dimensional linear space, the bilinear form* $B(\boldsymbol{x}, \boldsymbol{y})$ *can be uniquely represented in the form*

$$B(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i,\, j=1}^{n} b_{ij}\xi_i\eta_j \tag{7.3}$$

where

$$b_{ij} = B(\boldsymbol{e}_i, \boldsymbol{e}_j), \tag{7.4}$$

and ξ_i *and* η_j *are the coordinates of the vectors* $\boldsymbol{x}$ *and* $\boldsymbol{y}$, *respectively, in the basis* e.

Proof. Assume that $\boldsymbol{x} = \sum_{i=1}^{n} \xi_i \boldsymbol{e}_i$ and $\boldsymbol{y} = \sum_{j=1}^{n} \eta_j \boldsymbol{e}_j$ are the representations of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in terms of the basis e. The form $B(\boldsymbol{x}, \boldsymbol{y})$ being linear with respect to each of the arguments $\boldsymbol{x}$ and $\boldsymbol{y}$ (see (7.1.)), we have

$$B(\boldsymbol{x}, \boldsymbol{y}) = B\left(\sum_{i=1}^{n} \xi_i \boldsymbol{e}_i, \sum_{j=1}^{n} \eta_j \boldsymbol{e}_j\right) = \sum_{i,\, j=1}^{n} B(\boldsymbol{e}_i, \boldsymbol{e}_j)\, \xi_i \eta_j.$$

Thus, representation (7.3) with expressions (7.4) for the coefficients b_{ij} is valid for the form $B(\boldsymbol{x}, \boldsymbol{y})$.

To prove the uniqueness of this representation, we assume that representation (7.3) with **certain** coefficients b_{ij} is valid for $B(\boldsymbol{x}, \boldsymbol{y})$. Taking $\boldsymbol{x} = \boldsymbol{e}_i$, $\boldsymbol{y} = \boldsymbol{e}_j$ in (7.3), we immediately get expressions (7.4) for the coefficients b_{ij}. We have proved the theorem.

Definition. *The matrix*

$$B(e) = (b_{ij}) = \begin{pmatrix} b_{11} & b_{12} & & b_{1n} \\ b_{21} & b_{22} & & b_{2n} \\ \cdot & \cdot & & \cdot \\ b_{n1} & b_{n2} & & b_{nn} \end{pmatrix}, \tag{7.5}$$

whose elements b_{ij} are defined with the aid of relations (7.4) *is called the* ***matrix of the bilinear form*** $B(\boldsymbol{x}, \boldsymbol{y})$ *in the given basis* e.

Remark 1. Let us consider the question of constructing all the bilinear forms in a given finite-dimensional real space L. The answer to this question is the following: *any square matrix* (b_{ij}) *is a matrix of a certain bilinear form in the given basis* $e=(\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)$.

Let us verify the validity of this assertion.

Using the matrix (b_{ij}), we shall define the number function $B(\boldsymbol{x}, \boldsymbol{y})$ of two arguments $\boldsymbol{x}=\sum_{i=1}^{n} \xi_i \boldsymbol{e}_i$ and $\boldsymbol{y}=\sum_{j=1}^{n} \eta_j \boldsymbol{e}_j$ of the form $B(\boldsymbol{x}, \boldsymbol{y})=\sum_{i,\, j=1}^{n} b_{ij}\xi_i\eta_j$ in the linear space L with the given basis $e=(\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)$.

It is easy to see that this function satisfies all the conditions of the definition of a bilinear form. But then, according to Theorem 7.1, the elements b_{ij} of the given matrix are equal to $B(\boldsymbol{e}_i, \boldsymbol{e}_j)$, and the formula written above is the representation of that form as (7.3).

According to the remark made above, it is natural to call representation (7.3) of the bilinear form $B(\boldsymbol{x}, \boldsymbol{y})$ the **general form of a bilinear form in an n-dimensional linear space.**

Remark 2. If $B(\boldsymbol{x}, \boldsymbol{y})$ is a symmetric (skew-symmetric) bilinear form, then matrix (7.5) of that form in the basis e is symmetric (skew-symmetric). The converse is also true: if matrix (7.5) of the bilinear form $B(\boldsymbol{x}, \boldsymbol{y})$ is symmetric (skew-symmetric), then the bilinear form is symmetric (skew-symmetric) as well.

Let us verify the validity of this remark.

Assume that $B(\boldsymbol{x}, \boldsymbol{y})$ is a symmetric (skew-symmetric) bilinear form. Setting $\boldsymbol{x}=\boldsymbol{e}_i$, $\boldsymbol{y}=\boldsymbol{e}_j$ in (7.2), we find, in accordance with (7.4), that

$$b_{ij}=b_{ji}, \ (b_{ij}=-b_{ji}), \tag{7.6}$$

i.e. matrix (7.5) is symmetric (skew-symmetric).

Suppose now that matrix (7.5) of the bilinear form $B(\boldsymbol{x}, \boldsymbol{y})$ is symmetric (skew-symmetric), i.e. its elements satisfy relations (7.6). Then it follows from relation (7.3) and the relation $B(\boldsymbol{y}, \boldsymbol{x})=\sum_{i,\, j=1}^{n} b_{ji}\xi_i\eta_j$ that $B(\boldsymbol{x}, \boldsymbol{y})=B(\boldsymbol{y}, \boldsymbol{x})$, $(B(\boldsymbol{x}, \boldsymbol{y})=-B(\boldsymbol{y}, \boldsymbol{x}))$, i.e. the form $B(\boldsymbol{x}, \boldsymbol{y})$ is symmetric (skew-symmetric).

7.1.3. Transformation of a matrix of a bilinear form in passing to a new basis. The rank of a bilinear form. Let us consider two bases $e=(\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)$ and $f=(\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n)$ in the linear space L. Assume that $A(e)=(a_{ij})$ and $A(f)=(b_{ij})$ are the matrices of the given bilinear form in the indicated bases.

Let us consider the connection between these matrices, i.e. the transformation of the matrix a_{ij} of the bilinear form when we pass from the basis e to a new basis f.

The following statement holds true.

Theorem 7.2. *The matrices $A(e)$ and $A(f)$ of the bilinear form $A(x, y)$ in the bases $e=(e_1, e_2, \ldots, e_n)$ and $f=(f_1, f_2, \ldots, f_n)$ are related as*

$$A(f)=C'A(e)C, \tag{7.7}$$

where $C=(c_{pq})$ is the transition matrix from the basis e to the basis f, and C' is the transposed matrix C.

Proof. The elements f_q of the new basis f can be expressed in terms of the elements e_p of the old basis e with the aid of the matrix $C=(c_{pq})$ by the formulas

$$f_q=\sum_{p=1}^{n} c_{pq}e_p. \tag{7.8}$$

Since $b_{lk}=A(f_l, f_k)$, we get, in accordance with (7.8),

$$b_{lk}=A(f_l, f_k)=A\left(\sum_{i=1}^{n} c_{il}e_i, \sum_{j=1}^{n} c_{jk}e_j\right)$$
$$=\sum_{i,\,j=1}^{n} A(e_i, e_j)\,c_{il}c_{jk}=\sum_{i,\,j=1}^{n} a_{ij}c_{il}c_{jk}. \tag{7.9}$$

Recall that the elements c'_{li} of the transposed matrix C' are connected with the elements c_{il} of the matrix C by the relation $c_{il}=c'_{li}$.

Substituting these relations into the right-hand side of (7.9), we get the following expression for b_{lk};

$$b_{lk}=\sum_{i,\,j=1}^{n} a_{ij}c'_{li}c_{jk}=\sum_{i=1}^{n} c'_{li}\left(\sum_{j=1}^{n} a_{ij}c_{jk}\right). \tag{7.10}$$

The sum $\sum_{j=1}^{n} a_{ij}c_{jk}$ (by the definition of the product of matrices) is an element of the matrix $A(e)C$. Hence it follows that the expression on the right-hand side of (7.10) is an element of the matrix $C'A(e)C$. But the left-hand side of (7.10) includes an element of the matrix $A(f)$ and, therefore, $A(f)=C'A(e)C$. We have proved the theorem.

Corollary. *The rank of the matrix $A(f)$ is equal to that of the matrix $A(e)$.*

This follows immediately from (7.7), from the fact that the matrix C, and, hence, the matrix C' are nonsingular, and from the theorem stating that the rank of a matrix does not change when it is multiplied by a nonsingular matrix.

This corollary makes it possible to introduce an important **numerical invariant** of a bilinear form, the so-called rank of a bilinear form.

Definition 1. *The rank of a bilinear form defined in a finite-dimensional linear space L is known as the rank of the matrix of that form in an arbitrary basis of the space L.*

Definition 2. *The bilinear form $A(\boldsymbol{x}, \boldsymbol{y})$ given in the finite-dimensional linear space L is said to be nonsingular (singular) if its rank is equal to (smaller than) the dimension of the space L.*

7.2. Quadratic Forms

Assume that $A(\boldsymbol{x}, \boldsymbol{y})$ is a symmetric bilinear form defined on the linear space L.

Definition 1. *A quadratic form is a number function $A(\boldsymbol{x}, \boldsymbol{x})$ of the same vector argument $\boldsymbol{x}$ which is obtained from the bilinear form $A(\boldsymbol{x}, \boldsymbol{y})$ for $\boldsymbol{x}=\boldsymbol{y}$.*

The symmetric bilinear form $A(\boldsymbol{x}, \boldsymbol{y})$ is said to be the **polar of the quadratic form** $A(\boldsymbol{x}, \boldsymbol{x})$.

The polar bilinear form $A(\boldsymbol{x}, \boldsymbol{y})$ and the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ are related as

$$A(\boldsymbol{x}, \boldsymbol{y})=\frac{1}{2}[A(\boldsymbol{x}+\boldsymbol{y}, \boldsymbol{x}+\boldsymbol{y})-A(\boldsymbol{x}, \boldsymbol{x})-A(\boldsymbol{y}, \boldsymbol{y})],$$

which follows from the obvious equality

$$A(\boldsymbol{x}+\boldsymbol{y}, \boldsymbol{x}+\boldsymbol{y})=A(\boldsymbol{x}, \boldsymbol{x})+A(\boldsymbol{x}, \boldsymbol{y})+A(\boldsymbol{y}, \boldsymbol{x})+A(\boldsymbol{y}, \boldsymbol{y})$$

and the property of symmetry of the form $A(\boldsymbol{x}, \boldsymbol{y})$.

Suppose a symmetric bilinear form $A(\boldsymbol{x}, \boldsymbol{y})$ which is the polar of the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ is given in the finite-dimensional linear space L. Suppose, in addition, that a basis $e=(\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)$ is specified in L.

In accordance with Theorem 7.1, the form $A(\boldsymbol{x}, \boldsymbol{y})$ can be represented as (7.3)

$$A(\boldsymbol{x}, \boldsymbol{y})=\sum_{i, j=1}^{n} a_{ij}\xi_i\eta_j, \tag{7.3}$$

where ξ_i and η_j are the coordinates of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, in the basis e. By virtue of symmetry of $A(\boldsymbol{x}, \boldsymbol{y})$

$$a_{ij}=a_{ji} \tag{7.11}$$

(see Remark 2 in 7.1.2).

Setting $\boldsymbol{x}=\boldsymbol{y}$ (i.e. $\eta_j=\xi_j$) in (7.3), we get the following *represen-*

tation for the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ in the finite-dimensional space L with the given basis e:

$$A(\boldsymbol{x}, \boldsymbol{x}) = \sum_{i,\, j=1}^{n} a_{ij}\xi_i\xi_j. \tag{7.12}$$

The matrix (a_{ij}) is called a ***matrix of the quadratic form*** *$A(\boldsymbol{x}, \boldsymbol{x})$ in the given basis e.*

According to (7.11), the matrix (a_{ij}) is symmetric. Evidently, every symmetric matrix (a_{ij}) is associated with the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ with the aid of relation (7.12), relation (7.12) being a representation of $A(\boldsymbol{x}, \boldsymbol{x})$ in the space L with the given basis e (see also Remark 3 in 7.1.2).

Note that when we pass to a new basis, the matrix of a quadratic form is transformed by formula (7.7). Therefore, the rank of that matrix does not change when we pass to a new basis.

The *rank of the matrix of the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ is customarily called the* ***rank of a quadratic form***.

If the rank of the matrix of a quadratic form is equal to the dimension of the space L, the form is said to be **nonsingular,** otherwise it is **singular.**

In what follows, we shall use the following terminology.

Definition 2. *The quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ is called* (1) ***positive (negative) definite*** *if the inequality*

$$A(\boldsymbol{x}, \boldsymbol{x}) > 0 \quad (A(\boldsymbol{x}, \boldsymbol{x}) < 0)$$

is satisfied for any nonzero $\boldsymbol{x}$ (forms of this kind are also called ***definite****);*

(2) ***alternating*** *if there are $\boldsymbol{x}$ and $\boldsymbol{y}$ such that*

$$A(\boldsymbol{x}, \boldsymbol{x}) > 0, \quad A(\boldsymbol{y}, \boldsymbol{y}) < 0;$$

(3) ***quasi-definite*** *if*

$$A(\boldsymbol{x}, \boldsymbol{x}) \geqslant 0 \text{ or } A(\boldsymbol{x}, \boldsymbol{x}) \leqslant 0$$

for all $\boldsymbol{x}$, but there is a nonzero $\boldsymbol{x}$ for which

$$A(\boldsymbol{x}, \boldsymbol{x}) = 0.$$

We shall later indicate the criteria which can be used to decide to which of the indicated types the form $A(\boldsymbol{x}, \boldsymbol{x})$ belongs.

Note the following important ***assertion***.

If $A(\boldsymbol{x}, \boldsymbol{y})$ is a bilinear form which is the polar of the positive definite quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$, then $A(\boldsymbol{x}, \boldsymbol{y})$ satisfies all the axioms of the scalar product of vectors in a Euclidean space.

Let us direct our attention to the four axioms of a scalar product (see 4.1.1).

If the number called the scalar product of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is

symbolized as $A(\boldsymbol{x}, \boldsymbol{y})$, then the axioms can be written as follows:

1°. $A(\boldsymbol{x}, \boldsymbol{y}) = A(\boldsymbol{y}, \boldsymbol{x})$.

2°. $A(\boldsymbol{x}+\boldsymbol{z}, \boldsymbol{y}) = A(\boldsymbol{x}, \boldsymbol{y}) + A(\boldsymbol{z}, \boldsymbol{y})$.

3°. $A(\lambda\boldsymbol{x}, \boldsymbol{y}) = \lambda A(\boldsymbol{x}, \boldsymbol{y})$.

4°. $A(\boldsymbol{x}, \boldsymbol{x}) \geqslant 0$ and $A(\boldsymbol{x}, \boldsymbol{x}) > 0$ for $\boldsymbol{x} \neq 0$.

Since the bilinear form $A(\boldsymbol{x}, \boldsymbol{y})$, which is the polar of the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$, is symmetric, Axiom 1° is satisfied. Axioms 2° and 3°, in conjunction with the requirement of symmetry, are satisfied by virtue of the definition of a bilinear form (see 7.1.1). Axiom 4° is satisfied since the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ is positive definite.

Remark. The axioms of a scalar product can evidently be regarded as a collection of requirements defining a bilinear form which is the polar of a positive definite quadratic form. A scalar product in linear spaces can, therefore, be defined with the aid of a bilinear form of this kind.

7.3. Reducing a Quadratic Form to the Sum of Squares

We describe here various methods of reducing a quadratic form to the sum of squares, i.e. show the ways of choosing the basis $f = (\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n)$ in the linear space L relative to which the quadratic form can be represented in the following **canonical form:**

$$A(\boldsymbol{x}, \boldsymbol{x}) = \lambda_1\eta_1^2 + \lambda_2\eta_2^2 + \ldots + \lambda_n\eta_n^2, \tag{7.13}$$

$(\eta_1, \eta_2, \ldots, \eta_n)$ being the coordinates of $\boldsymbol{x}$ in the basis f.

The coefficients $\lambda_1, \lambda_2, \ldots, \lambda_n$ in expression (7.13) are called **canonical coefficients.**

It should be emphasized that we consider quadratic forms in an arbitrary real linear space. In 7.6 we shall study quadratic forms in a Euclidean space and shall prove the possibility of reducing every quadratic form to a canonical one even in an orthonormal basis. Proceeding from the results obtained in Chapter 5, we shall obtain, in 7.6, a new proof of the theorem on the reduction of a quadratic form to the canonical one in an arbitrary (not necessarily Euclidean) real linear space.

The present section is devoted not only to proving the possibility of reducing a quadratic form to the canonical form but also to a description of two methods of such a reduction which are very valuable in practical applications and are often encountered in practice.

Since every transformation of a basis is associated with a regular linear transformation of coordinates and a regular transformation of coordinates is associated with a transformation of a basis, the question concerning a reduction of a form to the canonical form can be decided by choosing a corresponding regular transformation of coordinates.

7.3.1. Lagrange's method. Let us prove the following theorem.

Theorem 7.3. *Any quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ defined in the n-dimensional linear space L can be reduced to the canonical form (7.13) by means of a regular linear transformation of coordinates.*

Proof. We present the proof of the theorem by **Lagrange's method.** This method consists in a successive completion of a quadratic trinomial with respect to every argument to form a perfect square.

We assume that $A(\boldsymbol{x}, \boldsymbol{x}) \not\equiv 0$* and in the given basis $e=(\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)$ has the form

$$A(\boldsymbol{x}, \boldsymbol{x}) = \sum_{i,\, j=1}^{n} a_{ij}\xi_i\xi_j. \tag{7.14}$$

We shall verify, first, that by means of a regular transformation of coordinates the form $A(\boldsymbol{x}, \boldsymbol{x})$ can be transformed so that the *coefficient in the square of the first coordinate of the vector x is nonzero.*

If in the given basis this coefficient is nonzero, then the required regular transformation is identical.

If $a_{11}=0$, but the coefficient in the square of some other coordinate is different from zero, then we can renumber the base vectors and attain the necessary result. It is clear that the renumbering is a regular transformation.

Now if all the coefficients in the squares of the coordinates are zero, the necessary transformation can be obtained as follows. Assume, for instance, $a_{12} \neq 0$**. Let us consider the following regular transformation of coordinatess***:

$$\xi_1' = \xi_1 - \xi_2,\ \xi_2' = \xi_1 + \xi_2,\ \xi_i' = \xi_i,\ i = 3, 4, \ldots, n.$$

After this transformation, the coefficient in $\xi_1'^2$ is equal to $2a_{12}$ and is, therefore, nonzero.

Thus, we assume that $a_{11} \neq 0$ in relation (7.14).

Let us isolate, in expression (7.14), the group of the terms containing ξ_1. We get

$$A(\boldsymbol{x}, \boldsymbol{x}) = a_{11}\xi_1^2 + 2a_{12}\xi_1\xi_2 + \ldots + 2a_{1n}\xi_1\xi_n + \sum_{i,\, j=2}^{n} a_{ij}\xi_i\xi_j. \tag{7.15}$$

---

* If $A(\boldsymbol{x}, \boldsymbol{x}) \equiv 0$, then its matrix in any basis consists of zero elements and, therefore, such a form has a canonical form in any basis by definition.

** Recall that $A(\boldsymbol{x}, \boldsymbol{x}) \not\equiv 0$ and, therefore, at least one of the coefficients a_{ij} is different from zero.

*** The determinant of the matrix of this transformation is equal to 2 and the transformation is, therefore, regular.

We transform the isolated group of the terms as follows.

$$a_{11}\xi_1^2+2a_{12}\xi_1\xi_2+\ldots+2a_{1n}\xi_1\xi_n$$

$$=a_{11}\left(\xi_1+\frac{a_{12}}{a_{11}}\xi_2+\ldots+\frac{a_{1n}}{a_{11}}\xi_n\right)^2$$

$$-\frac{a_{12}^2}{a_{11}}\xi_2^2-\ldots-\frac{a_{1n}^2}{a_{11}}\xi_n^2-2\frac{a_{12}a_{13}}{a_{11}}\xi_2\xi_3$$

$$-\ldots-2\frac{a_{1n-1}a_{1n}}{a_{11}}\xi_{n-1}1\xi_n.$$

It is evident that expression (7.15) can be rewritten as

$$A(\boldsymbol{x},\ \boldsymbol{x})=a_{11}\left(\xi_1+\frac{a_{12}}{a_{11}}\xi_2+\ldots+\frac{a_{1n}}{a_{11}}\xi_n\right)^2+\sum_{i,\ j=2}^{n}a^*_{ij}\xi_i\xi_j, \quad (7.16)$$

where a^*_{ij} are coefficients in $\xi_i\xi_j$ obtained as a result of the transformation.

Let us consider the following regular transformation of coordinates

$$\begin{aligned}
\eta_1&=\xi_1+\frac{a_{12}}{a_{11}}\xi_2+\ldots+\frac{a_{1n}}{a_{11}}\xi_n,\\
\eta_2&=\qquad\quad\xi_2,\\
&\cdot\qquad\qquad\qquad\cdot\ \cdot\ \cdot\\
\eta_n&=\qquad\qquad\qquad\qquad\xi_n.
\end{aligned}$$

By means of this transformation and representation (7.16) for $A(\boldsymbol{x},\ \boldsymbol{x})$ we get

$$A(\boldsymbol{x},\ \boldsymbol{x})=a_{11}\eta_1^2+\sum_{i,\ j=2}^{n}a^*_{ij}\eta_i\eta_j. \quad (7.17)$$

Thus, if the form $A(\boldsymbol{x},\ \boldsymbol{x})\not\equiv 0$, then by means of regular transformation of coordinates this form can be reduced to (7.17).

Let us consider now the quadratic form $\sum_{i,\ j=2}^{n}a^*_{ij}\eta_i\eta_j$. If this form is identically zero, then the problem of reduction of $A(\boldsymbol{x},\ \boldsymbol{x})$ to the canonical form is solved. Now if the form $\sum_{i,\ j=2}^{n}a^*_{ij}\eta_i\eta_j\not\equiv 0$, we can repeat the arguments, considering the transformation of the coordinates $\eta_2,\ \ldots,\ \eta_n$, similar to those discussed above, without changing the coordinate η_1. The transformation of coordinates η_1, $\eta_2,\ \ldots,\ \eta_n$ of this type is, evidently, regular.

It is clear that in a finite number of steps we reduce the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ to the canonical form (7.13).

Note that we can obtain the necessary transformation of the original coordinates $\xi_1, \xi_2, \ldots, \xi_n$ by multiplying the regular transformation obtained in the process of reasoning. We have thus proved the theorem.

Remark 1. The basis in which a quadratic form has a canonical form is called **canonical.** Note that a canonical basis is not uniquely defined.

Remark 2. If the form $A(\boldsymbol{x}, \boldsymbol{x})$ is reduced to canonical equation (7.13), then, in general, not all the canonical coefficients λ_i are nonzero. Leaving only nonzero coefficients λ_i in (7.13) and renumbering them, we get the following expression for $A(\boldsymbol{x}, \boldsymbol{x})$:

$$A(\boldsymbol{x}, \boldsymbol{x}) = \lambda_1 \eta_1^2 + \lambda_2 \eta_2^2 + \ldots + \lambda_k \eta_k^2. \tag{7.18}$$

It is clear that $k \leqslant n$. Since, by definition, the rank of a quadratic form is equal to the rank of its matrix in any basis, it follows from (7.18) and the condition $\lambda_i \neq 0$ for $i = 1, 2, \ldots, k$ that the rank of the form is k. Thus, *the number of nonzero canonical coefficients is equal to the rank of the quadratic form.*

7.3.2. Jacobi's method. Under some additional assumptions concerning the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$, we can indicate explicit formulas of transition from the given basis $e = (\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)$ to the canonical *basis* $f = (\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n)$ and formulas for the canonical coefficients λ_i:

We first introduce the concept of the **triangular transformation** of the base vectors.

The transformation of the base vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ is said to be triangular if it has the form

$$\left.\begin{aligned} \boldsymbol{f}_1 &= \boldsymbol{e}_1, \\ \boldsymbol{f}_2 &= \alpha_{21}\boldsymbol{e}_1 + \boldsymbol{e}_2, \\ \boldsymbol{f}_3 &= \alpha_{31}\boldsymbol{e}_1 + \alpha_{32}\boldsymbol{e}_2 + \boldsymbol{e}_3, \\ &\cdots\cdots\cdots \\ \boldsymbol{f}_n &= \alpha_{n1}\boldsymbol{e}_1 + \alpha_{n2}\boldsymbol{e}_2 + \ldots + \boldsymbol{e}_n. \end{aligned}\right\} \tag{7.19}$$

Remark. Since the determinant of the matrix of the triangular transformation (7.19) is nonzero (equal to 1), the vectors $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n$ form a basis.

We introduce into consideration the angular minors of the matrix $A(e) = (a_{ij})$ of the coefficients of the form $A(\boldsymbol{x}, \boldsymbol{x})$ in the basis e, designating them as $\Delta_1, \Delta_2, \ldots, \Delta_{n-1}$:

$$\Delta_1 = a_{11}, \quad \Delta_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \ldots, \quad \Delta_{n-1} = \begin{vmatrix} a_{11} & \ldots & a_{1,n-1} \\ \cdot & & \cdot \\ a_{n-1,1} & \ldots & a_{n-1,n-1} \end{vmatrix}. \tag{7.20}$$

The following statement holds true.

Theorem 7.4. *Let the minors $\Delta_1, \Delta_2, \ldots, \Delta_{n-1}$ of the matrix (a_{ij}) of the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ be nonzero. Then there is a unique triangular transformation of the base vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ by means of which the form $A(\boldsymbol{x}, \boldsymbol{x})$ can be reduced to the canonical form.*

Proof. Recall that the coefficients b_{ij} of the form $A(\boldsymbol{x}, \boldsymbol{x})$ in the basis $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n$ can be calculated by the formulas $b_{ij}=A(\boldsymbol{f}_i, \boldsymbol{f}_j)$.

If in the basis $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n$ $A(\boldsymbol{x}, \boldsymbol{x})$ has a canonical form, then $b_{ij}=0$ for $i \neq j$. Therefore, to prove the theorem it is sufficient to construct, by means of the triangular transformation (7.19), a basis $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n$ such that the following relations hold true:

$$A(\boldsymbol{f}_i, \boldsymbol{f}_j)=0 \text{ for } i \neq j, \text{ or, which is the same, for } i<j \tag{7.21}$$

(we must verify, of course, that the required transformation is unique).

If we direct our attention to formulas (7.19) for $\boldsymbol{f}_i$, then, using the linear property of the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ with respect to each argument, we can easily see that relations (7.21) are satisfied if the relations*

$$A(\boldsymbol{e}_1, \boldsymbol{f}_j)=0,\ A(\boldsymbol{e}_2, \boldsymbol{f}_j)=0,\ \ldots,\ A(\boldsymbol{e}_{j-1}, \boldsymbol{f}_j)=0, \qquad j=2, 3, \ldots, n \tag{7.22}$$

hold true.

Let us write out formulas (7.22). For that purpose, we substitute the expression

$$\boldsymbol{f}_j=\alpha_{j1}\boldsymbol{e}_1+\alpha_{j2}\boldsymbol{e}_2+\ldots+\alpha_{j,\,j-1}\boldsymbol{e}_{j-1}+\boldsymbol{e}_j \tag{7.23}$$

for $\boldsymbol{f}_j$ from relations (7.19) into the left-hand sides of these formulas. Using the linear property of $A(\boldsymbol{x}, \boldsymbol{x})$ with respect to each argument and the designation $A(\boldsymbol{e}_i, \boldsymbol{e}_j)=a_{ij}$, we obtain the following linear system of equations for the unknown coefficients α_{jk}:

$$\left.\begin{array}{l} \alpha_{j1}a_{11}+\alpha_{j2}a_{12}+\ldots+\alpha_{j,\,j-1}a_{1,\,j-1}+a_{1j}=0, \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ \alpha_{j1}a_{j-1,\,1}+\alpha_{j2}a_{j-1,\,2}+\ldots+\alpha_{j,\,j-1}a_{j-1,\,j-1}+a_{j-1,\,j}=0. \end{array}\right\} \tag{7.24}$$

The determinant of this system is equal to Δ_{j-1}. By the hypothesis, $\Delta_{j-1} \neq 0$. Consequently, system (7.24) has a *unique* solution. We can thus construct a unique triangular transformation of the base vectors by means of which $A(\boldsymbol{x}, \boldsymbol{x})$ can be reduced to the canonical form. We have proved the theorem.

We present formulas for calculating the coefficients α_{ji} of the required triangular transformation and the formulas for the canonical coefficients λ_j.

* It is easy to ascertain that relations (7.21) yield relations (7.22).

We use the symbol $\Delta_{j-1,i}$ for the minor of the matrix (a_{ij}) lying in the rows of the matrix labelled $1, 2, \ldots, j-1$ and the columns labelled $1, 2, \ldots, i-1, i+1, \ldots, j$. Then, using system (7.24) and Cramer's rule, we get the following expression for α_{ji}:

$$\alpha_{ji}=(-1)^{j+i}\frac{\Delta_{j-1,i}}{\Delta_{j-1}}. \tag{7.25}$$

Let us now calculate the canonical coefficients λ_j.

Since $\lambda_j=b_{jj}=A(\boldsymbol{f}_j, \boldsymbol{f}_j)$, expression (7.23) for $\boldsymbol{f}_j$ and formulas (7.22) yield

$$\begin{aligned}\lambda_j=A(\boldsymbol{f}_j, \boldsymbol{f}_j)&=A(\alpha_{j1}\boldsymbol{e}_1+\alpha_{j2}\boldsymbol{e}_2+\ldots+\alpha_{j,j-1}\boldsymbol{e}_{j-1}\\&\quad+\boldsymbol{e}_j, \boldsymbol{f}_j)=A(\boldsymbol{e}_j, \boldsymbol{f}_j)=A(\boldsymbol{e}_j, \alpha_{j1}\boldsymbol{e}_1+\alpha_{j2}\boldsymbol{e}_2\\&\quad+\ldots+\alpha_{j,j-1}\boldsymbol{e}_{j-1}+\boldsymbol{e}_j)=\alpha_{j1}a_{1j}+\alpha_{j2}a_{2j}\\&\quad+\ldots+\alpha_{j,j-1}a_{j-1,j}+a_{jj}.\end{aligned}$$

Substituting expression (7.25) for α_{ji} into the right-hand side of the last relation, we find that

$$\begin{aligned}\lambda_j=((-1)^{j+1}a_{1j}\Delta_{j-1,1}&+(-1)^{j+2}a_{2j}\Delta_{j-1,2}+\ldots\\&\ldots+(-1)^{2j-1}a_{j-1,j}\Delta_{j-1,j-1}+a_{jj}\Delta_{j-1})\Delta_{j-1}^{-1}.\end{aligned}$$

The numerator of the last relation is the sum of the products of the elements of the row labelled j in the determinant Δ_j by the algebraic adjuncts of those elements in the indicated determinant. Consequently, the numerator is equal to Δ_j. Therefore,

$$\lambda_j=\frac{\Delta_j}{\Delta_{j-1}},\quad j=2, 3, \ldots, n. \tag{7.26}$$

Since $\lambda_1=A(\boldsymbol{f}_1, \boldsymbol{f}_1)=A(\boldsymbol{e}_1, \boldsymbol{e}_1)=a_{11}=\Delta_1$, we get from this and from (7.26) the following formulas for the canonical coefficients:

$$\lambda_1=\Delta_1,\quad \lambda_2=\frac{\Delta_2}{\Delta_1},\quad \ldots,\quad \lambda_n=\frac{\Delta_n}{\Delta_{n-1}}. \tag{7.27}$$

7.4. The Law of Inertia of Quadratic Forms. Classification of Quadratic Forms

7.4.1. The law of inertia of quadratic forms. We have pointed out (see Remark 2 in 7.3.1) that the rank of a quadratic form is equal to the number of nonzero canonical coefficients. Thus, the number of nonzero canonical coefficients does not depend on the choice of a regular transformation by means of which the form $A(\boldsymbol{x}, \boldsymbol{x})$ can be reduced to the canonical form. Indeed, the number of positive and negative canonical coefficients does not change with any

method of reducing $A(\boldsymbol{x}, \boldsymbol{x})$ to the canonical form. This property is called the **law of inertia** of quadratic forms.

Some remarks are due before passing to the substantiation of the law of inertia.

Suppose in the basis $e=(\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)$ the form $A(\boldsymbol{x}, \boldsymbol{x})$ is defined by the matrix $A(e)=(a_{ij})$:

$$A(\boldsymbol{x}, \boldsymbol{x})=\sum_{i, j=1}^{n} a_{ij}\xi_i\xi_j, \tag{7.28}$$

where $\xi_1, \xi_2, \ldots, \xi_n$ are the coordinates of the vector $\boldsymbol{x}$ in the basis e. We assume that with the aid of regular transformation of coordinates, this form has been reduced to the canonical form

$$A(\boldsymbol{x}, \boldsymbol{x})=\lambda_1\mu_1^2+\lambda_2\mu_2^2+\ldots+\lambda_k\mu_k^2, \tag{7.29}$$

$\lambda_1, \lambda_2, \ldots, \lambda_k$ being nonzero canonical coefficients labelled so that the first q coefficients are positive and the following coefficients are negative:

$$\lambda_1>0,\ \lambda_2>0,\ \ldots,\ \lambda_q>0,\ \lambda_{q+1}<0,\ \ldots,\ \lambda_k<0.$$

Let us consider the following regular transformation of the coordinates μ_i*:

$$\left.\begin{aligned} &\eta_1=\frac{1}{\sqrt{\lambda_1}}\mu_1,\ \eta_2=\frac{1}{\sqrt{\lambda_2}}\mu_2, \ldots, \eta_q=\frac{1}{\sqrt{\lambda_q}}\mu_q,\\ &\eta_{q+1}=\frac{1}{\sqrt{-\lambda_{q+1}}}\mu_{q+1}, \ldots, \eta_k=\frac{1}{\sqrt{-\lambda_k}}\mu_k,\\ &\eta_{k+1}=\mu_{k+1}, \ldots, \eta_n=\mu_n. \end{aligned}\right\} \tag{7.30}$$

As a result of this transformation $A(\boldsymbol{x}, \boldsymbol{x})$ assumes the form

$$A(\boldsymbol{x}, \boldsymbol{x})=\eta_1^2+\eta_2^2+\ldots+\eta_q^2-\eta_{q+1}^2-\ldots-\eta_k^2, \tag{7.31}$$

which is called the *normal form* of the quadratic form.

Thus, by means of a certain regular transformation of the coordinates $\xi_1, \xi_2, \ldots, \xi_n$ of the vector $\boldsymbol{x}$ in the basis $e=(\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)$

$$\begin{gathered} \eta_i=\alpha_{i1}\xi_1+\alpha_{i2}\xi_2+\ldots+\alpha_{in}\xi_n,\\ i=1, 2, \ldots, n,\ \det(\alpha_{ij})\neq 0 \end{gathered} \tag{7.32}$$

(this transformation is the product of the transformations of ξ into μ

* It is easy to see that the determinant of this transformation is different from zero.

and μ into η formulas (7.30)), the quadratic form can be reduced to normal form (7.31).

Let us prove the following assertion.

Theorem 7.5 (the law of inertia of quadratic forms). *The number of terms with positive (negative) coefficients in the normal form of a quadratic form does not depend on the way of reducing the quadratic form to that form.*

Proof. Assume that by means of regular transformation of coordinates (7.32) the form $A(\boldsymbol{x}, \boldsymbol{x})$ has been reduced to the normal form (7.31) and by means of another regular transformation of coordinates it has been reduced to the normal form

$$A(\boldsymbol{x}, \boldsymbol{x}) = \zeta_1^2 + \zeta_2^2 + \ldots + \zeta_p^2 - \zeta_{p+1}^2 - \ldots - \zeta_k^2. \tag{7.33}$$

To prove the theorem, it is, evidently, sufficient to verify the validity of the equality $p = q$.

Suppose $p > q$. Let us verify that in this case there is a nonzero vector $\boldsymbol{x}$ such that with respect to the bases in which this form $A(\boldsymbol{x}, \boldsymbol{x})$ has the forms (7.31) and (7.33), the coordinates $\eta_1, \eta_2, \ldots, \eta_q$ and $\zeta_{p+1}, \ldots, \zeta_n$ of that vector are equal to zero:

$$\eta_1 = 0, \ \eta_2 = 0, \ldots, \ \eta_q = 0, \ \zeta_{p+1} = 0, \ldots, \ \zeta_n = 0. \tag{7.34}$$

Since the coordinates η_i have been obtained by means of regular transformation (7.32) of the coordinates $\xi_1, \ldots, \xi_n$, and the coordinates ξ_j have been obtained by means of a similar regular transformation of the same coordinates ξ_1, ξ_n, relations (7.34) can be regarded as a system of homogeneous linear equations with respect to the coordinates $\xi_1, \ldots, \xi_n$ of the required vector $\boldsymbol{x}$ in the basis $e = (\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)$ (for example, in accordance with (7.32), the relation $\eta_1 = 0$ can be written out as $\alpha_{11}\xi_1 + \alpha_{12}\xi_2 + \ldots + \alpha_{1n}\xi_n = 0$). Since $p > q$, the number of homogeneous equations (7.34) is less than n and, therefore, system (7.34) has a nonzero solution with respect to the coordinates $\xi_1, \ldots, \xi_n$ of the required vector $\boldsymbol{x}$. Consequently, if $p > q$, then there is a nonzero vector $\boldsymbol{x}$ such that relations (7.34) hold true.

Let us calculate the value of the form $A(\boldsymbol{x}, \boldsymbol{x})$ for the vector $\boldsymbol{x}$. Using relations (7.31) and (7.33), we get

$$A(\boldsymbol{x}, \boldsymbol{x}) = -\eta_{q+1}^2 - \ldots - \eta_k^2 = \zeta_1^2 + \zeta_2^2 + \ldots + \zeta_p^2.$$

The last equality holds only in the case $\eta_{q+1} = \ldots = \eta_k = 0$ and $\zeta_1 = \zeta_2 = \ldots = \zeta_p = 0$. Thus, in a certain basis all the coordinates $\zeta_1, \zeta_2, \ldots, \zeta_n$ of the nonzero vector $\boldsymbol{x}$ are equal to zero (see the last equations and relations (7.34)), i.e. the vector $\boldsymbol{x}$ is equal to zero. Consequently, the assumption $p > q$ leads to a contradiction. By similar considerations, $p < q$ leads to a contradiction too.

Thus, $p = q$. We have proved the theorem.

7.4.2. Classification of quadratic forms. In 7.2.1 (see Definition 2) we introduced the concepts of a positive definite, negative definite, alternating and quasi-definite quadratic forms.

In this subsection, with the aid of the concepts of inertia indices, of positive and negative indices of inertia of a quadratic form, we shall indicate the way of determining the type of a quadratic form. The *index of inertia* of a quadratic form is the number of nonzero canonical coefficients of that form (i.e. its rank), the *positive index of inertia* is the number of positive canonical coefficients and the *negative index of inertia* is the number of negative canonical coefficients. It is clear that the sum of a positive and a negative index of inertia is equal to the index of inertia.

Thus, assume that the index of inertia, the positive and the negative index of inertia of the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ are equal to k, p, and q respectively ($k=p+q$). We proved in 7.4.1 that in any canonical basis $f=(\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n)$ this form could be reduced to the following normal form:

$$A(\boldsymbol{x}, \boldsymbol{x})=\eta_1^2+\eta_2^2+\ldots+\eta_p^2-\eta_{p+1}^2-\ldots-\eta_k^2, \tag{7.35}$$

where $\eta_1, \eta_2, \ldots, \eta_n$ are the coordinates of the vector $\boldsymbol{x}$ in the basis f.

1°. *The necessary and sufficient condition for a quadratic form to be definite.*

The following statement holds true.

For the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ given in the n-dimensional linear space L to be definite, it is necessary and sufficient that either the positive index of inertia p or the negative index of inertia q should be equal to the dimension n of the space L.

In this case, if $p=n$, then the form is positive definite, if $q=n$, then the form is negative definite.

Proof. Since the cases of a positive definite form and a negative definite form are treated analogously, we shall give the proof of the statement for positive definite forms.

(1) Necessity. Let the form $A(\boldsymbol{x}, \boldsymbol{x})$ be positive definite. Then expression (7.35) assumes the form

$$A(\boldsymbol{x}, \boldsymbol{x})=\eta_1^2+\eta_2^2+\ldots+\eta_p^2.$$

If in this case $p<n$, then it follows from the last expression that for the nonzero vector $\boldsymbol{x}$ with the coordinates

$$\eta_1=0,\ \eta_2=0,\ \ldots,\ \eta_p=0,\ \eta_{p+1}\neq 0,\ \ldots,\ \eta_n\neq 0$$

the form $A(\boldsymbol{x}, \boldsymbol{x})$ vanishes, and this contradicts the definition of a positive definite quadratic form. Consequently, $p=n$.

(2) Sufficiency. Assume $p=n$. Then relation (7.35) has the form $A(\boldsymbol{x}, \boldsymbol{x})=\eta_1^2+\eta_2^2+\ldots+\eta_n^2$. It is clear that $A(\boldsymbol{x}, \boldsymbol{x})\geqslant 0$, and if $A=0$, then $\eta_1=\eta_2=\ldots=\eta_n=0$, i.e. the vector $\boldsymbol{x}$ is zero. Consequently, $A(\boldsymbol{x}, \boldsymbol{x})$ is a positive definite form.

Remark. To use the indicated criterion to decide whether a quadratic form is definite, we must reduce the quadratic form to the canonical form.

In 7.4.3 we shall prove *Sylvester's criterion* for the definiteness of a quadratic form which can be used to decide whether a form given in any basis without being reduced to a canonical form is definite.

2°. *The necessary and sufficient condition for a quadratic form to be alternating.*

Let us prove the following statement.

For a quadratic form to be alternating, it is necessary and sufficient that both the positive and the negative indices of inertia of that form be nonzero.

Proof. **(1) Necessity.** Since an alternating form can assume positive as well as negative values, its representation (7.35) in the normal form must contain both positive and negative terms (otherwise, the form would assume either nonnegative or nonpositive values). Consequently, both positive and negative indices of inertia are different from zero.

(2) Sufficiency. Assume $p \neq 0$ and $q \neq 0$. Then we have $A(\boldsymbol{x}_1, \boldsymbol{x}_1) > 0$ for the vector $\boldsymbol{x}_1$ with the coordinates $\eta_1 \neq 0, \ldots, \eta_p \neq 0, \eta_{p+1} = 0, \ldots, \eta_n = 0$ and $A(\boldsymbol{x}_2, \boldsymbol{x}_2) < 0$ for the vector $\boldsymbol{x}_2$ with the coordinates $\eta_1 = 0, \ldots, \eta_p = 0, \eta_{p+1} \neq 0, \ldots, \eta_n \neq 0$. Consequently, $A(\boldsymbol{x}, \boldsymbol{x})$ is an alternating form.

3°. *The necessary and sufficient condition for a quadratic form to be quasi-definite.*

The following statement holds true.

For the form $A(\boldsymbol{x}, \boldsymbol{x})$ to be quasi-definite, it is necessary and sufficient that the following relations hold true: either $p < n$, $q = 0$ or $p = 0$, $q < n$.

Proof. We shall consider the case of a positive quasi-definite form. The case of a negative quasi-definite form can be treated by analogy.

(1) Necessity. Let the form $A(\boldsymbol{x}, \boldsymbol{x})$ be positive quasi-definite. Then, evidently, $q = 0$ and $p < n$ (in the case $p = n$ the form would be positive definite).

(2) Sufficiency. If $p < n$, $q = 0$, then $A(\boldsymbol{x}, \boldsymbol{x}) \geqslant 0$ and we have $A(\boldsymbol{x}, \boldsymbol{x}) = 0$ for the nonzero vector $\boldsymbol{x}$ with the coordinates $\eta_1 = 0, \ldots, \eta_p = 0, \eta_{p+1} \neq 0, \ldots, \eta_n \neq 0$, i.e. $A(\boldsymbol{x}, \boldsymbol{x})$ is a positive quasi-definite form.

7.4.3. Sylvester's* criterion for the definiteness of a quadratic form. Assume that the form $A(\boldsymbol{x}, \boldsymbol{x})$ in the basis $e = (\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)$ is defined by the matrix $A(e) = (a_{ij})$: $A(\boldsymbol{x}, \boldsymbol{x}) = \sum_{i,j=1}^{n} a_{ij}\xi_i\xi_j$,

* Sylvester, James Joseph (1814-1897), an English mathematician.

and that $\Delta_1 = a_{11}$, $\Delta_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \ldots, \Delta_n = \begin{vmatrix} a_{11} & a_{1n} \\ a_{n1} & a_{nn} \end{vmatrix}$ are angular minors and the determinant of the matrix (a_{ij}). The following statement holds true.

Theorem 7.6 (Sylvester's criterion). *For the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ to be positive definite, it is necessary and sufficient that the inequalities $\Delta_1 > 0$, $\Delta_2 > 0, \ldots, \Delta_n > 0$ be satisfied.*

For a quadratic form to be negative definite, it is necessary and sufficient that the signs of the angular minors alternate, and $\Delta_1 < 0$.

Proof. **(1) Necessity.** We shall first prove that the condition of definiteness of the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ yields $\Delta_i \neq 0$, $i = 1, 2, \ldots, n$.

Let us verify that the assumption $\Delta_k = 0$ leads to a contradiction, that under this assumption there is a nonzero vector $\boldsymbol{x}$ such that $A(\boldsymbol{x}, \boldsymbol{x}) = 0$, which contradicts the definiteness of the form.

Thus, assume $\Delta_k = 0$. Let us consider the following quadratic homogeneous system of linear equations:

$$\left.\begin{aligned} a_{11}\xi_1 + a_{12}\xi_2 + \ldots + a_{1k}\xi_k &= 0, \\ a_{21}\xi_1 + a_{22}\xi_2 + \ldots + a_{2k}\xi_k &= 0, \\ \cdots\cdots\cdots\cdots & \\ a_{k1}\xi_1 + a_{k2}\xi_2 + \ldots + a_{kk}\xi_k &= 0. \end{aligned}\right\} \tag{7.36}$$

Since Δ_k is a determinant of this system and $\Delta_k = 0$, system (7.36) has a nonzero solution $\xi_1, \xi_2, \ldots, \xi_k$ (not all ξ_i are zero). Let us multiply the first equation (7.36) by ξ_1, the second equation by $\xi_2, \ldots,$ the last equation by ξ_k and add the resulting relations together. As a result we obtain an equality $\sum_{i,\, j=1}^{k} a_{ij}\xi_i\xi_j = 0$ whose left-hand side is a value of the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ for the nonzero vector $\boldsymbol{x}$ with the coordinates $(\xi_1, \xi_2, \ldots, \xi_k, 0, \ldots, 0)$. This value is equal to zero which contradicts the definiteness of the system.

We have thus verified that $\Delta_i \neq 0$, $i = 1, 2, \ldots, n$. We can, therefore, apply Jacobi's method for reducing the form $A(\boldsymbol{x}, \boldsymbol{x})$ to the sum of squares (see Theorem 7.4) and use formulas (7.27) for the canonical coefficients λ_i. If $A(\boldsymbol{x}, \boldsymbol{x})$ is a positive definite form, then all the canonical coefficients are positive. But then it follows from relations (7.27) that $\Delta_1 > 0$, $\Delta_2 > 0, \ldots, \Delta_n > 0$. Now if $A(\boldsymbol{x}, \boldsymbol{x})$ is a negative definite form, then all the canonical coefficients are negative. But then it follows from formulas (7.27) that the signs of the angular minors alternate, and $\Delta_1 < 0$.

(2) Sufficiency. Suppose the conditions imposed on the angular minors Δ_i in the formulation of the theorem are fulfilled. Since

$\Delta_i \neq 0$, $i=1, 2, \ldots, n$, the form A can be reduced to the sum of squares by Jacobi's method (see Theorem 7.4), and the canonical coefficients λ_i can be found by formulas (7.27). If $\Delta_1 > 0, \Delta_2 > 0, \ldots$ $\ldots, \Delta_n > 0$, then it follows from relations (7.27) that all $\lambda_i > 0$, i.e. the form $A(\boldsymbol{x}, \boldsymbol{x})$ is positive definite. Now if the signs of Δ_i alternate and $\Delta_1 < 0$, then it follows from relations (7.27) that the form $A(\boldsymbol{x}, \boldsymbol{x})$ is negative definite. And this is what we wished to prove.

7.5. Multilinear Forms

Definition. *The multilinear form $A(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p)$ of p vector arguments is a number function which is defined on various vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ of the linear space L and is linear with respect to each argument, the values of the other arguments being fixed.*

The product of the linear forms $A(\boldsymbol{x}_1) A(\boldsymbol{x}_2) \ldots A(\boldsymbol{x}_p)$ can serve as a simple example of a multilinear form.

The multilinear form $A(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p)$ is said to be *symmetric (skew-symmetric)* if for every two of its arguments $\boldsymbol{x}_k$ and $\boldsymbol{x}_l$ and for any values of these arguments the following relation holds true:

$$A(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k, \ldots, \boldsymbol{x}_l, \ldots, \boldsymbol{x}_p) = A(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l, \ldots, \boldsymbol{x}_k, \ldots, \boldsymbol{x}_p)$$

$$(A(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k, \ldots, \boldsymbol{x}_l, \ldots, \boldsymbol{x}_p) = -A(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l, \ldots, \boldsymbol{x}_k, \ldots, \boldsymbol{x}_p)).$$

Suppose the multilinear form $A(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p)$ is defined in the finite-dimensional linear space L and let $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ be a basis in L. Let us resolve each vector $\boldsymbol{x}_i$ into the components $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$:

$$\boldsymbol{x}_i = \xi_{i1}\boldsymbol{e}_1 + \xi_{i2}\boldsymbol{e}_2 + \ldots + \xi_{in}\boldsymbol{e}_n = \sum_{j=1}^{n} \xi_{ij}\boldsymbol{e}_j, \quad i=1, 2, \ldots, p. \tag{7.37}$$

Substituting the expressions for $\boldsymbol{x}_i$ from formulas (7.37) into the multilinear form $A(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p)$ and using the property of linearity of that form with respect to each argument, we get

$$A(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p) = A\Big(\sum_{j_1=1}^{n} \xi_{1j_1}\boldsymbol{e}_{j_1}, \sum_{j_2=1}^{n} \xi_{2j_2}\boldsymbol{e}_{j_2}, \ldots, \sum_{j_p=1}^{n} \xi_{pj_p}\boldsymbol{e}_{j_p}\Big)$$

$$= \sum_{j_1, j_2, \ldots, j_p=1}^{n} \xi_{1j_1}\xi_{2j_2} \ldots \xi_{pj_p} A(\boldsymbol{e}_{j_1}, \boldsymbol{e}_{j_2}, \ldots, \boldsymbol{e}_{j_p}). \tag{7.38}$$

Thus the values of the multilinear form $A(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p)$ in a finite-dimensional space with the isolated basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ are defined by various values $A(\boldsymbol{e}_{j_1}, \boldsymbol{e}_{j_2}, \ldots, \boldsymbol{e}_{j_p})$ of that form on the vectors $\boldsymbol{e}_{j_1}, \boldsymbol{e}_{j_2}, \ldots, \boldsymbol{e}_{j_p}$.

Let us prove the following assertion.

Theorem 7.7. *Any multilinear skew-symmetric form $A(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ defined in the n-dimensional linear space L with the isolated basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ can be represented as*

$$A(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) = a \begin{vmatrix} \xi_{11} & \xi_{12} & & \xi_{1n} \\ \xi_{21} & \xi_{22} & & \xi_{2n} \\ \cdot & \cdot & & \\ \xi_{n1} & \xi_{n2} & & \xi_{nn} \end{vmatrix}, \tag{7.39}$$

where $a = A(\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n)$, $a(\xi_{i1}, \xi_{i2}, \ldots, \xi_{in})$ are the coordinates of the vectors $\boldsymbol{x}_i$ in the basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$.

Proof. Since the form $A(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ is skew-symmetric to perform an arbitrary permutation $(j_1, j_2, \ldots, j_n)$ of the indices $(1, 2, \ldots, n)$ we have

$$\begin{aligned} A(\boldsymbol{e}_{j_1}, \boldsymbol{e}_{j_2}, \ldots, \boldsymbol{e}_{j_n}) &= (-1)^{N(j_1, j_2, \ldots j_n)} A(\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n) \\ &= (-1)^{N(j_1, j_2, \ldots, j_n)} a, \end{aligned} \tag{7.40}$$

where $N(j_1, j_2, \ldots, j_n)$ is the number of disorders in the permutation $(j_1, j_2, \ldots, j_n)$.

Because of the skew-symmetry of the form, the value $A(\boldsymbol{e}_{j_1}, \ldots, \boldsymbol{e}_{j_k}, \ldots, \boldsymbol{e}_{j_l}, \ldots, \boldsymbol{e}_{j_n})$ is equal to zero for two identical indices j_k and j_l $(j_k = j_l)$. It follows from this and from (7.40) that for the case being considered relation (7.38) assumes the form

$$\begin{aligned} &A(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) \\ &= a \sum_{j_1, j_2, \ldots, j_n = 1}^{n} (-1)^{N(j_1, j_2, \ldots, j_n)} \xi_{1j_1}, \xi_{2j_2}, \ldots, \xi_{nj_n}. \end{aligned} \tag{7.41}$$

Comparing formula (7.41) with formula (1.28) in Chapter 1 for a determinant of order n, we ascertain the validity of relation (7.39). We have proved the theorem.

7.6. Bilinear and Quadratic Forms in a Euclidean Space

In previous sections we studied bilinear and quadratic forms in an arbitrary (not necessarily Euclidean) real linear space L. Here we shall obtain data on bilinear and quadratic forms defined in a real Euclidean space. We shall widely use the results obtained in 5.9 devoted to linear operators.

It will be shown in 7.6.3 how the theory of Euclidean spaces can be applied to obtain comprehensive results in arbitrary linear spaces. In particular, we shall obtain an independent proof of the theorem sta-

ting that every quadratic form in a linear space can be reduced to a canonical form.

7.6.1. Preliminary remarks. We shall recall here some notions of the theory of linear operators.

Let V be a n-dimensional real Euclidean space and $\boldsymbol{A}$, a linear operator from V into V. The operator A^* is said to be an adjoint of $\boldsymbol{A}$ if the equality

$$(\boldsymbol{Ax}, \boldsymbol{y})=(\boldsymbol{x}, \boldsymbol{A}^*\boldsymbol{y}) \tag{7.42}$$

is valid for all $\boldsymbol{x}\in V$ and $\boldsymbol{y}\in V$.

The operator $\boldsymbol{A}$ is said to be self-adjoint if $\boldsymbol{A}=\boldsymbol{A}^*$, i.e. if

$$(\boldsymbol{Ax}, \boldsymbol{y})=(\boldsymbol{x}, \boldsymbol{Ay}) \tag{7.43}$$

for all $\boldsymbol{x}\in V$ and $\boldsymbol{y}\in$V.

Let us consider a bilinear form $B(\boldsymbol{x}, \boldsymbol{y})$ defined in the Euclidean space V. We have found (see Chapter 5) that every such form $B(\boldsymbol{x}, \boldsymbol{y})$ is in a one-to-one correspondence with a linear operator such that the equality

$$B(\boldsymbol{x}, \boldsymbol{y})=(\boldsymbol{Ax}, \boldsymbol{y}) \tag{7.44}$$

holds true.

In addition, it has been proved (Theorem 5.33) that the bilinear form $B(\boldsymbol{x}, \boldsymbol{y})$ is symmetric if and only if the operator $\boldsymbol{A}$ appearing in (7.44) is self-adjoint.

It should also be pointed out that it has been proved (Theorem 5.35) that any self-adjoint operator $\boldsymbol{A}$ has an orthonormal basis of eigenvectors. This means that there exists an orthonomal system $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ and real numbers $\lambda_1, \lambda_2, \ldots, \lambda_n$ such that

$$\boldsymbol{Ae}_k=\lambda_k\boldsymbol{e}_k. \tag{7.45}$$

Note that in the basis $\{\boldsymbol{e}_k\}$ the matrix of the operator $\boldsymbol{A}$ has a diagonal form.

7.6.2. Reducing a quadratic form to the sum of squares in an orthogonal basis. Suppose $B(\boldsymbol{x}, \boldsymbol{y})$ is a symmetric bilinear form defined in the real Euclidean space V, and $B(\boldsymbol{x}, \boldsymbol{x})$ is a quadratic form it defines.

Let us prove the following theorem on reducing the quadratic form $B(\boldsymbol{x}, \boldsymbol{x})$ to the sum of squares.

Theorem 7.8. *Let $B(\boldsymbol{x}, \boldsymbol{y})$ be a symmetric bilinear form defined in the Euclidean space V. Then there exists, in the space V, an orthonormal basis $\{\boldsymbol{e}_k\}$ and real numbers such that for any $\boldsymbol{x}\in V$ the quadratic form $B(\boldsymbol{x}, \boldsymbol{x})$ can be represented as the following sum of the squares of the coordinates ξ_k of the vector $\boldsymbol{x}$ in the basis $\{\boldsymbol{e}_k\}$:*

$$B(\boldsymbol{x}, \boldsymbol{x})=\sum_{k=1}^{n} \lambda_k\xi_k^2. \tag{7.46}$$

Proof. Since $B(\boldsymbol{x}, \boldsymbol{y})$ is a symmetric bilinear form, there is a self-adjoint operator $\boldsymbol{A}$ such that

$$B(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}). \tag{7.47}$$

By Theorem 5.35, we can indicate, for the operator $\boldsymbol{A}$, an orthonormal basis $\{\boldsymbol{e}_k\}$ of the eigenvectors of the operator; let λ_k be the eigenvalues corresponding to $\boldsymbol{e}_k$.

Assume that the vector $\boldsymbol{x}$ has coordinates ξ_k in the basis $\{\boldsymbol{e}_k\}$:

$$\boldsymbol{x} = \sum_{k=1}^{n} \xi_k \boldsymbol{e}_k. \tag{7.48}$$

Then, since $\boldsymbol{e}_k$ are the eigenvectors of the operator $\boldsymbol{A}$, we evidently have

$$\boldsymbol{A}\boldsymbol{x} = \sum_{k=1}^{n} \lambda_k \xi_k \boldsymbol{e}_k. \tag{7.49}$$

Since the basis $\{\boldsymbol{e}_k\}$ is orthonormal, we obtain the following expression for the scalar product $(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{x})$ from relations (7.48) and (7.49):

$$(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}) = \sum_{k=1}^{n} \lambda_k \xi_k^2. \tag{7.50}$$

From this and from relation (7.47) we get (7.46). We have proved the theorem.

7.6.3. Simultaneous reduction of two quadratic forms to the sum of squares in a linear space.

We shall now prove an important theorem on simultaneous reduction of two quadratic forms to the sum of squares in an arbitrary (not necessarily Euclidean) real linear space.

Theorem 7.9. *Assume that $A(\boldsymbol{x}, \boldsymbol{y})$ and $B(\boldsymbol{x}, \boldsymbol{y})$ are symmetric bilinear forms defined in the real linear space V. Assume furthermore that the inequality $B(\boldsymbol{x}, \boldsymbol{x}) > 0$ is valid for all $\boldsymbol{x} \in V$, $\boldsymbol{x} \neq 0$, (i.e. the quadratic form $B(\boldsymbol{x}, \boldsymbol{x})$ is positive definite). Then in the space V there is a basis $\{\boldsymbol{e}_k\}$ such that the quadratic forms $A(\boldsymbol{x}, \boldsymbol{x})$ and $B(\boldsymbol{x}, \boldsymbol{x})$ can be represented as*

$$A(\boldsymbol{x}, \boldsymbol{x}) = \sum_{k=1}^{n} \lambda_k \xi_k^2, \tag{7.51}$$

$$B(\boldsymbol{x}, \boldsymbol{x}) = \sum_{k=1}^{n} \xi_k^2 \quad , \tag{7.52}$$

where ξ_k are the coordinates of the vector $\boldsymbol{x}$ in the basis $\{\boldsymbol{e}_k\}$.

Proof. In accordance with the remark made at the end of 7.2, in a finite-dimensional real space a scalar product can be defined by means of the bilinear form $B(\boldsymbol{x}, \boldsymbol{y})$ which is the polar of the positive definite quadratic form $B(\boldsymbol{x}, \boldsymbol{x})$.

We can, therefore, introduce a scalar product $(\boldsymbol{x}, \boldsymbol{y})$ of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in the linear space V setting

$$(\boldsymbol{x}, \boldsymbol{y}) = B(\boldsymbol{x}, \boldsymbol{y}). \tag{7.53}$$

Thus V is a Euclidean space with the scalar product (7.53). By Theorem 7.11, we can indicate an orthonormal basis $\{\boldsymbol{e}_k\}$ and real numbers λ_k such that the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ is in the form (7.51) in that basis.

On the other hand, in any orthonormal basis, the scalar product $(\boldsymbol{x}, \boldsymbol{x})$ which is equal, according to (7.53), to $B(\boldsymbol{x}, \boldsymbol{x})$, can be represented as the sum of the squares of the coordinates of the vector $\boldsymbol{x}$. Thus we have also substantiated the representation $B(\boldsymbol{x}, \boldsymbol{x})$ in the form (7.52) in the basis $\{\boldsymbol{e}_k\}$. And this is what we wished to prove.

Remark. It follows directly from the theorem we have proved that any quadratic form in an arbitrary real linear space can be reduced to a canonical form. However, the method of this reduction is, in general, more complicated than the methods presented in 7.3 since it requires finding all the eigenvectors of a self-adjoint operator (see Chapter 6).

7.6.4. Extremal properties of a quadratic form. Let us consider an arbitrary differentiable function f defined on a certain smooth surface S (see the definition of a smooth surface in Chapter 5 of our *Fundamentals of Mathematical Analysis,* Part 2). We say that the point $\boldsymbol{x}_0$ of the surface S is a **stationary point** of the function f if at the point $\boldsymbol{x}_0$ the derivative of the function f in any direction on the surface S is equal to zero. In particular, the points of extremum of the function f are its stationary points.

The value $f(\boldsymbol{x}_0)$ of the function f at the stationary point $\boldsymbol{x}_0$ is called its **stationary value.** The stationary point $\boldsymbol{x}_0$ of the function f is sometimes called its **critical point,** and the value, $f(\boldsymbol{x}_0)$ its **critical value.** We shall discuss here stationary and, in particular, extremal values of the quadratic form $B(\boldsymbol{x}, \boldsymbol{x})$ on a sphere of unit radius in the Euclidean space V and show how these values are connected with the eigenvalues of the self-adjoint operator $\boldsymbol{A}$ by means of which the symmetric bilinear form $B(\boldsymbol{x}, \boldsymbol{y})$, which is the polar of the quadratic form $B(\boldsymbol{x}, \boldsymbol{x})$, can be represented as

$$B(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}). \tag{7.54}$$

A *unit sphere in* V is the set of vectors $\boldsymbol{x} \in V$ which satisfy the equation

$$(\boldsymbol{x}, \boldsymbol{x}) = 1 \quad \text{or} \quad \|\boldsymbol{x}\| = 1. \tag{7.55}$$

To simplify the arguments, we shall use the conclusions made in 7.6.3 concerning the reduction of a quadratic form to the sum of squares.

Thus, suppose $B(\boldsymbol{x}, \boldsymbol{x})$ is a quadratic form, $B(\boldsymbol{x}, \boldsymbol{y})$ is a bilinear form which is the polar of that form and $\boldsymbol{A}$ is a self-adjoint operator related to $B(\boldsymbol{x}, \boldsymbol{y})$ as (7.54).

By Theorem 7.8, in the orthonormal basis $\{\boldsymbol{e}_k\}$ consisting of the eigenvectors of the operator $\boldsymbol{A}$ the quadratic form $B(\boldsymbol{x}, \boldsymbol{x})$ has the form

$$B(\boldsymbol{x}, \boldsymbol{x}) = \sum_{k=1}^{n} \lambda_k \xi_k^2, \tag{7.56}$$

where ξ_k are the coordinates of the vector $\boldsymbol{x}$ in the basis $\{\boldsymbol{e}_k\}$, and λ_k are the eigenvalues of the operator $\boldsymbol{A}$. We agree to enumerate these eigenvalues in a decreasing order:

$$\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n. \tag{7.57}$$

Note that in the chosen basis the unit sphere, defined by equation (7.55), in the coordinates of the vector $\boldsymbol{x}$ is specified by the equation

$$\sum_{k=1}^{n} \xi_k^2 - 1 = 0. \tag{7.58}$$

Let us prove the following theorem.

Theorem 7.10. *The stationary values of the quadratic form $B(\boldsymbol{x}, \boldsymbol{x})$ on unit sphere (7.55) are equal to the eigenvalues λ_k of the operator $\boldsymbol{A}$. These stationary values can be attained, in particular, on the unit eigenvectors $\boldsymbol{e}_k$ of the operator $\boldsymbol{A}$.*

Proof. Since we speak of the *stationary values of the function $B(\boldsymbol{x}, \boldsymbol{x})$ under the condition $(\boldsymbol{x}, \boldsymbol{x}) = 1$, i.e. of the conditional extremum* of that function, we can apply Lagrange's method of undetermined multipliers (see our *Fundamentals of Mathematical Analysis,* Part 1, 15.5.2). Let us form Lagrange's function $\Psi(\xi_1, \xi_2, \ldots, \xi_n)$ for the function $B(\boldsymbol{x}, \boldsymbol{x})$ using its expression (7.56) in the given basis $\{\boldsymbol{e}_k\}$ and taking into account that the connection relation has the form (7.58). We get

$$\Psi = \sum_{k=1}^{n} \lambda_k \xi_k^2 - \lambda \left(\sum_{k=1}^{n} \xi_k^2 - 1 \right), \tag{7.59}$$

where λ is Lagrange's undetermined multiplier.

Recall that if we choose λ in (7.59) such that under condition (7.58) the relations

$$\frac{\partial \Psi}{\partial \xi_k} = 0, \quad k = 1, 2, \ldots, n \tag{7.60}$$

hold true, then at the points of the sphere (7.58), which correspond to those values of λ, the function $B(\boldsymbol{x}, \boldsymbol{x})$ (the quadratic form $B(\boldsymbol{x}, \boldsymbol{x})$) has a stationary value.

Thus the problem concerning the stationary values of $B(\boldsymbol{x}, \boldsymbol{x})$ on the sphere $(\boldsymbol{x}, \boldsymbol{x})=1$ reduces to the investigation of the behaviour of the system of equations (7.58), (7.60) with respect to the variables λ and the coordinates $\xi_1, \xi_2, \ldots, \xi_n$ of the vector $\boldsymbol{x}$. Note that in this case $\xi_1, \xi_2, \ldots, \xi_n$ are the coordinates of the vector $\boldsymbol{x}$ on which $B(\boldsymbol{x}, \boldsymbol{x})$ has a stationary value.

Since $\frac{\partial \Psi}{\partial \xi_k}=2(\lambda_k-\lambda)\xi_k$, system (7.58), (7.60) in question assumes the form

$$\left.\begin{aligned} &\sum_{k=1}^{n} \xi_k^2=1, \\ &(\lambda_k-\lambda)\xi_k=0, \quad k=1, 2, \ldots, n. \end{aligned}\right\} \tag{7.61}$$

Let system (7.61) have a solution

$$\lambda=\tilde{\lambda}, \quad \tilde{\boldsymbol{x}}=(\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_n).$$

Multiplying each relation $(\lambda_k-\tilde{\lambda})\tilde{\xi}_k=0$ by $\tilde{\xi}_k$, we then add the resulting relations together and, taking into account that $\sum_{k=1}^{n} \tilde{\xi}_k^2=1$, we get, according to (7.56) the following value for $\tilde{\lambda}$:

$$\tilde{\lambda}=\sum_{k=1}^{n} \lambda_k \tilde{\xi}_k^2=B(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}}).$$

Thus *if* $\tilde{\lambda}$ *and* $\tilde{\boldsymbol{x}}=(\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_n)$ *is a solution of system* (7.61), *then* $\tilde{\lambda}$ *is equal to the value of the quadratic form* $B(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}})$ *on the vector* $\tilde{\boldsymbol{x}}=(\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_n)$ *on which this form has a stationary value.*

It is easy to see that the following values of the unknowns λ and ξ_i are solutions of system (7.61):

$$\lambda=\lambda_k, \quad \xi_1=0, \ldots, \xi_{k-1}=0, \quad \xi_k=1, \quad \xi_{k+1}=0, \ldots, \xi_n=0,$$
$$k=1, 2, \ldots, n.$$

These solutions are, evidently, the eigenvalues λ_k and the coordinates of the corresponding eigenvectors $\boldsymbol{e}_k$. We have proved the theorem.

Remark. We have just found out that the eigenvalues λ_k are the stationary values of the quadratic form $B(\boldsymbol{x}, \boldsymbol{x})$ on the sphere $(\boldsymbol{x}, \boldsymbol{x})=1$.

It turns out that the numbers λ_1 and λ_n (under the condition (7.57)) are the greatest and the least value of $B(\boldsymbol{x}, \boldsymbol{x})$, respectively, on the sphere $(\boldsymbol{x}, \boldsymbol{x})=1$ (we have established above that these values can be attained).

To verify the validity of the remark, it is sufficient to replace all λ_k in (7.56) first by λ_n and then by λ_1 and use relations (7.57) and (7.58). Then we evidently obtain inequalities $\lambda_n \leqslant B(\boldsymbol{x}, \boldsymbol{x}) \leqslant \lambda_1$.

7.7. Second-Order Hypersurface

In this section we shall get acquainted with the main types of second-order hypersurfaces and indicate the means for investigating those surfaces.

7.7.1. The concept of a second-order hypersurface. Let V be an n-dimensional real Euclidean space.

For the sake of geometrical vividness, we shall call the vectors $\boldsymbol{x}$ of that space its **points.**

A second-order ***hypersurface*** *(hyperquadric) S is a locus of points $\boldsymbol{x}$ satisfying an equation of the form*

$$A(\boldsymbol{x}, \boldsymbol{x})+2B(\boldsymbol{x})+c=0, \tag{7.62}$$

where $A(\boldsymbol{x}, \boldsymbol{x})$ is a quadratic form which is not identically zero, $B(\boldsymbol{x})$ is a linear form and c is a real number.

Equation (7.62) is a **general equation of a second-order hypersurface.**

We isolate an orthonormal basis $\{\boldsymbol{e}_k\}$ in the space V, and designate the coordinates of the vector $\boldsymbol{x}$ (point $\boldsymbol{x}$) in that basis as $(x_1, x_2, \ldots, x_n)$. Then (see 7.1.2) the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ can be represented as

$$A(\boldsymbol{x}, \boldsymbol{x})=\sum_{j,\ k=1}^{n} a_{jk}x_jx_k, \tag{7.63}$$

where

$$a_{jk}=A(\boldsymbol{e}_j, \boldsymbol{e}_k), \tag{7.64}$$

and $A(\boldsymbol{e}_j, \boldsymbol{e}_k)$ is the value of the symmetric bilinear form $A(\boldsymbol{x}, \boldsymbol{y})$ which is the polar of the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$, on the vectors $\boldsymbol{e}_j$ and $\boldsymbol{e}_k$.

In the indicated basis $\{\boldsymbol{e}_k\}$ the linear form $B(\boldsymbol{x})$ can be represented in the form

$$B(\boldsymbol{x})=\sum_{k=1}^{n} b_kx_k. \tag{7.65}$$

Thus *the general equation of a second-order hypersurface in the*

Euclidean space V with the isolated basis $\{e_k\}$ can be represented in the form

$$\sum_{j,\ k=1}^{n} a_{jk}x_jx_k+2\sum_{k=1}^{n} b_kx_k+c=0. \tag{7.66}$$

Let us agree on the following terminology.

We shall call the term $A(\boldsymbol{x},\ \boldsymbol{x})=\sum\limits_{j,\ k=1}^{n} a_{jk}x_jx_k$ the **group of the leading terms** of equation (7.62) or (7.66).

The group of terms $B(\boldsymbol{x})+c=\sum\limits_{k=1}^{n} b_kx_k+c$ will be called the **linear part** of equation (7.62) or (7.66).

In what follows we shall consider the matrices

$$A=\begin{pmatrix} a_{11} & & a_{1n} \\ \cdot & \cdot & \cdot \\ a_{n1} & & a_{nn} \end{pmatrix} \text{ and } B=\begin{pmatrix} a_{11} & & a_{1n} & b_1 \\ & & & \\ a_{n1} & & a_{nn} & b_n \\ b_1 & \dots & b_n & c \end{pmatrix} \tag{7.67}$$

and the determinants det A and det B of those matrices.

The investigation of second-order hypersurfaces will be carried out by the method similar to that used in analytic geometry in investigating curves and surfaces of second order defined by general equations.

The method consists in the fact that by choosing a special Cartesian system of coordinates on a plane (for second-order curves) or in space (for second-order surfaces) the equation of a curve or a surface is simplified as far as possible. Then, in the process of the investigation of that equation the geometrical properties of the curve or the surface are elucidated. In addition, the enumeration of all possible types of the elementary equations of second-order curves and surfaces enables us to classify them.

To use this method in a multi-dimensional case, we must first investigate the transformations (mappings) of an n-dimensional Euclidean space which are analogs of the transformations of rectangular Cartesian coordinates in the case of two or three dimensions.

In an n-dimensional case, transformations of this kind are translations and the transformations of bases which map an orthonormal basis into a new orthonormal basis. Exact definitions of these transformations will be given in 7.7.2.

A second-order hypersurface treated as a geometrical object of the space V does not, evidently, change when transformations of the kind indicated above are performed. We shall later ascertain that

for every equation of the form (7.62) (or (7.66)), we can choose an origin and an orthonormal basis in V such that this equation written in the coordinates with respect to the new basis will be very simple and, therefore, as in the case of two or three dimensions, we can indicate the geometrical characteristics of such surfaces and classify them.

7.7.2. Translations in a Euclidean space. Transformations of orthonormal bases into orthonormal bases. A translation or **parallel displacement** in the Euclidean space V is a transformation defined by the relation

$$\boldsymbol{x}=\boldsymbol{x}'+\mathring{\boldsymbol{x}}, \tag{7.68}$$

where $\mathring{\boldsymbol{x}}$ is a fixed point called the origin of coordinates, or simply origin.

Let the points $\boldsymbol{x}$, $\boldsymbol{x}'$ and $\mathring{\boldsymbol{x}}$ possess coordinates equal respectively to $(x_1, x_2, \ldots, x_n)$, $(x'_1, x'_2, \ldots, x'_n)$, and $(\mathring{x}_1, \mathring{x}_2, \ldots, \mathring{x}_n)$.

Then, in coordinates, the translation is defined by the formulas

$$x_k=x'_k+\mathring{x}_k, \quad k=1, 2, \ldots, n. \tag{7.69}$$

Note that a translation does not change any fixed basis.

Let us now elucidate the characteristics of the transformation of an orthonormal basis into an orthonormal basis.

We assume that the orthonormal basis $\{\boldsymbol{e}_k\}$ is transformed into a new orthonormal basis $\{\boldsymbol{e}'_k\}$. Let us express every vector $\boldsymbol{e}'_k$ in terms of the vectors $\{\boldsymbol{e}_k\}$. We obtain

$$\left.\begin{aligned} \boldsymbol{e}'_1&=p_{11}\boldsymbol{e}_1+p_{21}\boldsymbol{e}_2+\ldots+p_{n1}\boldsymbol{e}_n,\\ \boldsymbol{e}'_2&=p_{12}\boldsymbol{e}_1+p_{22}\boldsymbol{e}_2+\ldots+p_{n2}\boldsymbol{e}_n,\\ &\ldots\ldots\ldots\ldots\ldots\ldots\\ \boldsymbol{e}'_n&=p_{1n}\boldsymbol{e}_1+p_{2n}\boldsymbol{e}_2+\ldots+p_{nn}\boldsymbol{e}_n. \end{aligned}\right\} \tag{7.70}$$

We designate the matrix of transformation (7.70) as P:

$$P=\begin{pmatrix} p_{11} & p_{21} & & p_{n1}\\ p_{12} & p_{22} & & p_{n2}\\ \cdot & \cdot & & \cdot\\ p_{1n} & p_{2n} & & p_{nn} \end{pmatrix} \tag{7.71}$$

Since the basis $\{\boldsymbol{e}_k\}$ and $\{\boldsymbol{e}'_k\}$ are orthonormal, we get from (7.70), by means of a scalar multiplication of $\boldsymbol{e}'_j$ by $\boldsymbol{e}'_k$,

$$(\boldsymbol{e}'_j, \boldsymbol{e}'_k)=\sum_{m=1}^{n} p_{mj}p_{mk}=\delta_{jk}=\begin{cases}1 & \text{for } j=k\\ 0 & \text{for } j\neq k.\end{cases} \tag{7.72}$$

Let us now consider the transposed matrix P', i.e. the matrix obtained from P by interchanging rows and columns.

It is evident, according to (7.72), that

$$PP' = P'P = I, \tag{7.73}$$

where I is an identity matrix.

Equations (7.73) show that the matrix P' is an inverse of P, i.e.

$$P^{-1} = P'. \tag{7.74}$$

Let us now assume that we are considering a transformation of the orthonormal basis $\{e_k\}$ by formulas (7.70) and that the matrix P of that transformation satisfies condition (7.73) (or, which is the same, (7.74)).

Then, evidently, the elements p_{jk} of the matrix P satisfy condition (7.72), which is equivalent to the condition of orthonormality of the basis $\{e'_k\}$ in accordance with the same relations (7.72).

Recall that in 5.9 we called the marix P satisfying conditions (7.73) orthogonal.

Thus, *for transformation (7.70) to be a transformation of an orthonormal basis into an orthonormal basis, it is necessary and sufficient that the matrix P of that transformation be orthogonal.*

Remark. Directing our attention to formulas (5.14) for the transformation of the coordinates of a vector in the transformation of a basis (see 5.2.1) and taking into consideration that the inverse matrix of the orthogonal matrix P is a matrix P', we obtain the following formulas for the transformation of the coordinates of the point x when passing from one orthonormal basis to another:

$$\left.\begin{aligned} x_1 &= p_{11}x'_1 + p_{12}x'_2 + \ldots + p_{1n}x'_n, \\ x_2 &= p_{21}x'_1 + p_{22}x'_2 + \ldots + p_{2n}x'_n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n &= p_{n1}x'_1 + p_{n2}x'_2 + \ldots + p_{nn}x'_n. \end{aligned}\right\} \tag{7.75}$$

7.7.3. Transformation of a general equation of a second-order hypersurface in translation. Let us consider a translation which is defined as transformation of the space V by formula (7.68) (or in coordinates by formula (7.69)).

After replacing x by its expression from formula (7.68) by virtue of the linearity of a quadratic form with respect to the first and the second argument* and the properties of linear form, the left-hand

* The quadratic form $A(x, x)$ is connected with the symmetric bilinear form $A(x, y)$ which is a polar of the form $A(x, x)$. The bilinear form $A(x, y)$ is linear with respect to the arguments x and y. The expression $A(x', \mathring{x})$ appearing in what follows is the value of the form $A(x, y)$ on the vectors x' and $\mathring{x}$.

side of (7.62) assumes the following form:

$$A(\boldsymbol{x}', \boldsymbol{x}')+2[A(\boldsymbol{x}', \mathring{\boldsymbol{x}})+B(\boldsymbol{x}')]=[A(\mathring{\boldsymbol{x}}, \mathring{\boldsymbol{x}})+2B(\mathring{\boldsymbol{x}})+c]=0.$$

Thus, in translation (7.68), the general equation (7.62) for the hypersurface S will be written in the form

$$A(\boldsymbol{x}', \boldsymbol{x}')+2B'(\boldsymbol{x}')+c'=0, \tag{7.76}$$

where the linear form $B'(\boldsymbol{x}')$ and the constant number c' are defined by the relations

$$B'(\boldsymbol{x}')=A(\boldsymbol{x}', \mathring{\boldsymbol{x}})+B(\boldsymbol{x}'), \tag{7.77}$$

$$c'=A(\mathring{\boldsymbol{x}}, \mathring{\boldsymbol{x}})+2B(\mathring{\boldsymbol{x}})+c. \tag{7.78}$$

Let us write the resulting formulas in coordinates.

Suppose the coordinates of the points $\boldsymbol{x}'$ and $\mathring{\boldsymbol{x}}$ are equal to $x_1', x_2', \ldots, x_n'$ and $\mathring{x}_1, \mathring{x}_2, \ldots, \mathring{x}_n$ respectively. Since the basis $\{\boldsymbol{e}_k\}$ does not change in translation, the quadratic form $A(\boldsymbol{x}', \boldsymbol{x}')$ can be written as follows:

$$A(\boldsymbol{x}', \boldsymbol{x}')=\sum_{j,\,k=1}^{n} a_{jk}x_j'x_k' \tag{7.79}$$

(note that the coefficients $a_{jk}=A(\boldsymbol{e}_j, \boldsymbol{e}_k)$ do not change since the base vectors $\boldsymbol{e}_k$ remain unchanged).

Consequently, we can make an important conclusion: *in translation the group of the leading terms retains its form.*

Let us now consider formulas (7.77) and (7.78). Since

$$A(\boldsymbol{x}', \mathring{\boldsymbol{x}})=\sum_{k=1}^{n}\left(\sum_{j=1}^{n} a_{jk}\mathring{x}_j\right)x_k',$$

$$B(\boldsymbol{x}')=\sum_{k=1}^{n} b_k x_k',$$

$$A(\mathring{\boldsymbol{x}}, \mathring{\boldsymbol{x}})=\sum_{j,\,k=1}^{n} a_{jk}\mathring{x}_j\mathring{x}_k,$$

$$B(\mathring{\boldsymbol{x}})=\sum_{k=1}^{n} b_k\mathring{x}_k,$$

formula (7.77), assumes the form

$$B'(\boldsymbol{x}')=\sum_{k=1}^{n} b_k'x_k'=\sum_{k=1}^{n}\left[\sum_{j=1}^{n} a_{jk}\mathring{x}_j+b_k\right]x_k', \tag{7.80}$$

and formula (7.78) is written as

$$c'=\sum_{j,\,k=1}^{n} a_{jk}\mathring{x}_j\mathring{x}_k+2\sum_{k=1}^{n} b_k\mathring{x}_k+c. \tag{7.81}$$

Thus, equation (7.76) in coordinates has the form

$$\sum_{j,\,k=1}^{n} a_{jk}x_j'x_k' + 2\sum_{k=1}^{n} b_k'x_k' + c' = 0. \tag{7.82}$$

We shall need an expression for c' different from (7.81). Let us write (7.81) as follows:

$$c' = \sum_{k=1}^{n}\left[\sum_{j=1}^{n} a_{jk}\mathring{x}_j + b_k\right]\mathring{x}_k + \sum_{k=1}^{n} b_k\mathring{x}_k + c. \tag{7.83}$$

Keeping in mind that the coefficients b_k' are expressed, as follows from (7.80), by the formulas

$$b_k' = \sum_{j=1}^{n} a_{jk}\mathring{x}_j + b_k, \tag{7.84}$$

we get from (7.83) the required expression for c':

$$c' = \sum_{k=1}^{n} (b_k' + b_k)\mathring{x}_k + c. \tag{7.85}$$

7.7.4. Transformation of a general equation for a second-order hypersurface when passing from an orthonormal basis to an orthonormal one. Suppose the orthonormal basis $\{e_k\}$ is transformed into a new orthonormal basis $\{e_k'\}$ by formulas (7.70), and P is an orthogonal matrix of that transformation (see (7.71)). Then, in accordance with Remark 2 of this section, the coordinates x_k and x_k' of a point in the bases $\{e_k\}$ and $\{e_k'\}$ are related as (7.75). Substituting the expression for x_k from (7.75) into the left-hand side of equation (7.66) and taking into account that because of the homogeneity of relations (7.75) the group of the leading terms and the linear part of equation (7.66) are transformed autonomously we get the following expression for the general equation of a second-order hypersurface in the coordinates x_k' of the points in the transformed basis $\{e_k'\}$:

$$\sum_{j,\,k=1}^{n} a_{jk}'x_j'x_k' + 2\sum_{k=1}^{n} b_k'x_k' + c' = 0. \tag{7.86}$$

In accord with the autonomy of transformation of the group of the leading terms, there hold equalities

$$\left.\begin{aligned} \sum_{j,\,k=1}^{n} a_{jk}'x_j'x_k' &= \sum_{j,\,k=1}^{n} a_{jk}x_jx_k, \\ \sum_{k=1}^{n} b_k'x_k' &= \sum_{k=1}^{n} b_kx_k, \\ c' &= c. \end{aligned}\right\} \tag{7.87}$$

Directing our attention to the first formula (7.87), we see that to determine the coefficients a'_{jk}, we can use the rule for the transformation of the coefficients of quadratic form in passing to a new basis. Namely, if we use the designation A' for the matrix of the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ in the basis $\{\boldsymbol{e}'_k\}$, then, according to Theorem 7.2 and the relation $P'=P^{-1}$, *we get the following relation between the matrices A and A' of the form $A(\boldsymbol{x}, \boldsymbol{x})$ in the bases $\{\boldsymbol{e}_k\}$ and $\{\boldsymbol{e}'_k\}$:*

$$A'=P^{-1}AP \tag{7.88}$$

(recall that P is a matrix of an orthogonal transformation).

We shall now regard the matrix A' as the matrix of a certain linear operator $\boldsymbol{A}$ in the basis $\{\boldsymbol{e}'_k\}$*, and the matrix P^{-1} as the matrix of the passage from the basis $\{\boldsymbol{e}'_k\}$ to the basis $\{\boldsymbol{e}_k\}$. Then, according to Theorem 5.7 (see 5.2.2), the matrix A can be considered to be the matrix of the linear operator $\boldsymbol{A}$ in the basis $\{\boldsymbol{e}_k\}$.

In other words, *when we transform an orthonormal basis into an orthonormal one, the matrix of a quadratic form changes as the matrix of a certain linear operator.*

This inference will be used in 7.7.5.

Remark. Note that the operator $\boldsymbol{A}$ whose matrix in an orthonormal basis coincides with the matrix of the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ is *self-adjoint.*

To prove this statement, we present the following arguments.

Let $A(\boldsymbol{x}, \boldsymbol{x})$ be a quadratic form and $A(\boldsymbol{x}, \boldsymbol{y})$, a symmetric bilinear form which is the polar of $A(\boldsymbol{x}, \boldsymbol{x})$. According to Theorem 7.8, the bilinear form $A(\boldsymbol{x}, \boldsymbol{y})$ can be represented as

$$A(\boldsymbol{x}, \boldsymbol{y})=(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}),$$

where $\boldsymbol{A}$ is a *self-adjoint operator.*

The quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ can, therefore, be represented as

$$A(\boldsymbol{x}, \boldsymbol{x})=(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}).$$

Let us prove that in the orthonormal basis $\{\boldsymbol{e}_k\}$ the matrices of the operator $\boldsymbol{A}$ and of the quadratic form coincide. That will be the proof of the statement of the remark.

Suppose a_{jk} are the elements of the matrix of the form $A(\boldsymbol{x}, \boldsymbol{x})$ and $\tilde{a}_{jk}$ are the elements of the matrix of the operator $\boldsymbol{A}$ in the basis $\{\boldsymbol{e}_k\}$. According to 7.1.2, $a_{jk}=A(\boldsymbol{e}_j, \boldsymbol{e}_k)$ and, according to 5.2.1, formula (5.13), the elements $\tilde{a}_{jk}$ can be found from the equations $\boldsymbol{A}\boldsymbol{e}_j=\sum_{p=1}^{n} \tilde{a}_{jp}\boldsymbol{e}_p$.

* According to Theorem 5.5 (see 5.2.1), any square matrix consisting of n rows and n columns can be regarded as the matrix of a certain linear operator acting in an n-dimensional space.

We multiply scalarly both sides of the last relation by e_k. Then, taking the orthonormality of the basis $\{e_k\}$ into account, we get $(\tilde{A}e_j, e_k)=\tilde{a}_{jk}$. Since $A(e_j, e_k)=(Ae_j, e_k)$, we have $a_{jk}=\tilde{a}_{jk}$. We have proved the statement of the remark.

7.7.5. Invariants of the general equation of a second-order hypersurface. The **invariant** of general equation (7.62) (or (7.66)) of a second-order hypersurface relative to translations and transformations of the orthogonal bases into orthogonal ones is the function $f(a_{11}, a_{12}, \ldots, a_{nn}, b_1, \ldots, b_n, c)$ of the coefficients of this equation whose value does not change under the indicated transformations of a space.

Let us prove the following statement.

Theorem 7.11. *The invariants of general equation* (7.62) (or (7.66)) *of a second-order hypersurface are the coefficients of the characteristic polynomial of the matrix A of the quadratic form $A(x, x)$ and the determinant* $\det B$ *of the matrix B in relation* (7.67). *In particular, the invariants are* $\det A$ *and the trace* $a_{11}+a_{22}+\ldots a_{nn}$ *of the matrix A.*

Proof. It is evidently sufficient to prove the invariance of the quantities enumerated in the hypothesis separately for the translation and transformation of an orthonormal basis into an orthonormal one.

Let us first consider the *translation.* We established in 7.3 that the group of the leading terms retained its form under this transformation (see formula (7.79)). Therefore, the matrix A does not change and, consequently, neither does the characteristic polynomial of that matrix.

Let us prove the invariance of $\det B$.

As a result of the translation (7.68) (or (7.69)), the matrix is transformed into the matrix B', whose determinant, according to (7.82), has the form

$$\det B'=\begin{vmatrix} a_{11} & \ldots & a_{1n} & b'_1 \\ \cdot & & \cdot & \cdot \\ a_{n1} & \ldots & a_{nn} & b'_n \\ b'_1 & \ldots & b'_n & c' \end{vmatrix}, \tag{7.89}$$

where b'_k and c' can be determined from (7.84) and (7.85).

Let us subtract from the last $(n+1)$th row of the determinant (7.89) the elements of the first row multiplied by $\mathring{x}_1$, then the elements of the second row multiplied by $\mathring{x}_2$, and so on, and finally the elements of the nth row multiplied by $\mathring{x}_n$. Since the determinant remains unchanged under these transformaions, we can use (7.84) and

(7.85) and obtain a relation

$$\det B' = \begin{vmatrix} a_{11} & \dots & a_{1n}b'_n \\ & \cdot & \cdot \\ a_{n1} & \dots & a_{nn}b'_n \\ b_1 & & b'_n\left(\sum_{k=1}^{n} b_k\mathring{x}_k + c\right) \end{vmatrix} \tag{7.90}$$

Let us now subtract from the elements of the last $(n+1)$th column of determinant (7.90) the elements of the first column multiplied by $\mathring{x}_1$, then the elements of the second column multiplied by $\mathring{x}_2$, and so on, and, finally, the elements of the nth column multiplied by $\mathring{x}_n$. Since the determinant does not change under these transformations, we can use the relation $a_{jk}=a_{kj}$ following from the symmetry of the form $A(\boldsymbol{x}, \boldsymbol{y})$ and formula (7.84) and obtain $\det B$. Thus we have proved the equality $\det B' = \det B$. Consequently, $\det B$ is invariant relative to translations.

Let us now consider the *transformation of an orthonormal basis into an orthonormal one.*

We shall first verify that the coefficients of the characteristic polynomial of the matrix A of the quadratic form are the invariants of the transformation being considered.

We established in 7.7.4 that when we pass to a new orthonormal basis, the matrix A changes as the matrix of a certain linear operator. But in that case, as follows from Remark 1 of 5.2.3, the coefficients of the characteristic polynomial of that matrix do not change in passing to a new basis.

In particular, the determinant $\det A$ and the trace $a_{11}+a_{22}+\ldots+a_{nn}$ of the matrix A are invariants as the coefficients of a characteristic polynomial.

It remains to prove the invariance of the determinant $\det B$ in the transformation of an orthonormal basis into an orthonormal one.

Let us prove this fact.

We apply the following technique. We introduce the designations $b_k = a_{k,\,n+1}$, $k=1, 2, \ldots, n$, $c = a_{n+1,\,n+1}$. Then equation (7.66) of a hypersurface can be written as follows:

$$\sum_{j,\,k=1}^{n+1} a_{jk}x_jx_k = 0, \tag{7.91}$$

where $x_{n+1}=1$.

Let us consider the transformation of the variables $x_1, x_2, \ldots, x_n, x_{n+1}$ into the variables $x'_1, x'_2, \ldots, x'_n, x'_{n+1}$ under which the first n variables are transformed by formulas (7.75) and the variable x_{n+1} is transformed by the formula $x_{n+1}=x'_{n+1}$.

It is clear that this transformation of variables can be regarded

Chapter 8

TENSORS

In this chapter we consider important objects called **tensors** which are characterized, in every basis, by a collection of coordinates transformed, in a special way, in the passage from one basis to another. Tensors are widely employed in geometry, physics and mechanics. The concept of a tensor arises when we study various anisotropic phenomena (say, when we study the dependence of the distribution of the velocities of light propagation in a crystal on the direction of its propagation).

8.1. Transformation of Bases and Coordinates

This section dealing with the laws of transformation of coordinates in an arbitrary real Euclidean space E^n is of an auxiliary nature. The leading considerations arising in the process make the notion of a tensor introduced in the next section more visual.

8.1.1. Gram's* determinants. We shall indicate here the means of elucidating the linear dependence of the system of vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_k$ in a Euclidean space.

For that purpose, we introduce the so-called Gram's determinant, also called a Gramian, of the indicated system of vectors.

The **Gramian** of the system of vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_k$ is the following determinant:

$$\begin{vmatrix} (\boldsymbol{e}_1, \boldsymbol{e}_1) & (\boldsymbol{e}_1, \boldsymbol{e}_2) & & (\boldsymbol{e}_1, \boldsymbol{e}_k) \\ (\boldsymbol{e}_2, \boldsymbol{e}_1) & (\boldsymbol{e}_2, \boldsymbol{e}_2) & & (\boldsymbol{e}_2, \boldsymbol{e}_k) \\ \cdot & & & \\ (\boldsymbol{e}_k, \boldsymbol{e}_1) & (\boldsymbol{e}_k, \boldsymbol{e}_2) & & (\boldsymbol{e}_k, \boldsymbol{e}_k) \end{vmatrix} . \tag{8.1}$$

The following statement holds true.

Theorem 8.1. *For the system of vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_k$ of the Euclidean space E^n to be linearly dependent, it is necessary and sufficient that Gramian (8.1) of that system be zero.*

* Gram, Jorgen Pedersen (1850-1916), a Danish mathematician.

Let the point $\boldsymbol{x}$ with the coordinates $(x'_1, x'_2, \ldots, x'_n)$ lie on S. This means that its coordinates satisfy equation (7.94). The point $-\boldsymbol{x}$ with the coordinates $(-x'_1, -x'_2, \ldots, -x'_n)$ which is symmetric with respect to the point $\boldsymbol{x}$ about the point $\mathring{\boldsymbol{x}}$, evidently also lies on S since its coordinates also satisfy equation (7.94).

Thus, if the hypersurface S has a centre, then the *points of S are located symmetrically pairwise about the centre.*

Remark 1. If the second-order hypersurface S has a centre, then the invariants $\det A$, $\det B$ and the constant term c' in equation (7.94) are related as

$$\det B = c' \det A. \tag{7.95}$$

Indeed, we get for equation (7.94)

$$\det B = \begin{vmatrix} a_{11} & a_{1n} & 0 \\ \cdot & \cdot & \cdot \\ a_{n1} & a_{nn} & 0 \\ 0 & 0 & c' \end{vmatrix}.$$

The last formula yields (7.95).

The existence of a centre of a second-order hypersurface is connected with the solvability of the equations of centre (7.93).

If the equations of a centre have a unique solution, then the hypersurface S is called a ***central hypersurface.***

Since the determinant of system (7.93) is equal to $\det A$, and the necessary and sufficient condition for a unique solution of that system to exist is that its determinant should differ from zero, we can infer that *for the hypersurface S to be central, it is necessary and sufficient that* $\det A \neq 0$.

Remark 2. If the origin is transferred to the centre of the central hypersurface S, then the equation of that hypersurface has the form

$$\sum_{j,\,k=1}^{n} a_{jk} x_j x_k + \frac{\det B}{\det A} = 0. \tag{7.96}$$

Indeed, with the origin transferred to the centre, the equation of the hypersurface assumes the form (7.94). Since $\det A \neq 0$ for the central hypersurface, we find from (7.95) that $c' = \det B / \det A$. Substituting this expression for c' into (7.94), we obtain equation (7.96).

7.7.7. The standard simplification of any equation of a second-order hypersurface by transforming an orthonormal basis. By Theorem 7.8, there is an orthonormal basis in which the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ can be written as the sum of squares. Let us designate this basis as $\{\boldsymbol{e}'_k\}$ and the coordinates of the point $\boldsymbol{x}$ in that basis as $x'_1, x'_2, \ldots, x'_n$. Besides that, we use the letters $\lambda_1, \lambda_2, \ldots, \lambda_n$ to designate the eigenvalues of the self-adjoint operator $\boldsymbol{A}$ whose matrix in

the orthonormal basis coincides with that of the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ (see the remark in 7.7.4).

Using now the conclusions of Theorem 7.8, we write the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ in the coordinates $x'_1, x'_2, \ldots, x'_n$ of the point $\boldsymbol{x}$ in the basis $\{\boldsymbol{e}'_k\}$ as follows:

$$A(\boldsymbol{x}, \boldsymbol{x}) = \sum_{k=1}^{n} \lambda_k^2 x_k'^2. \tag{7.97}$$

Let us pass from the basis $\{\boldsymbol{e}_k\}$ to the basis $\{\boldsymbol{e}'_k\}$. Since the formulas for the transformation of the coordinates of points are linear and homogeneous under such a transformation (see the remark in 7.7.2, formulas (7.75)), the group of the leading terms and the linear part of the equation of the hypersurface S are transformed autonomously. By virtue of this and equation (7.97), the equation of the hypersurface S in the basis $\{\boldsymbol{e}'_k\}$ has the form*:

$$\sum_{k=1}^{n} \lambda_k x_k'^2 + 2\sum_{k=1}^{n} b'_k x'_k + c = 0. \tag{7.98}$$

The reduction of any equation of the second-order hypersurface S to the form (7.98) is called the ***standard simplification*** *of that equation (by the transformation of the orthonormal basis).*

7.7.8. Simplification of the equation of a central second-order hypersurface. Classification of central hypersurfaces. The conclusions made in 7.7.6 and 7.7.7 enable us to solve the problem of classifying all the central second-order hypersurfaces. The solution of the problem will be carried out by the following scheme. First, by transferring the origin to the centre of the hypersurface (7.66), we reduce the equation to the form (7.96). Then we perform a standard simplification of equation (7.96). As a result, we evidently get according to (7.98) the following equation of the central second-order hypersurface:

$$\lambda_1 x_1''^2 + \lambda_2 x_2''^2 + \ldots + \lambda_n x_n''^2 + \frac{\det B}{\det A} = 0, \tag{7.99}$$

in which λ_k are the eigennumbers of the matrix A of the quadratic form $A(\boldsymbol{x}, \boldsymbol{x})$ in equation (7.62), and x''_k are the coordinates of the point $\boldsymbol{x}$ in the final orthonormal basis $\{\boldsymbol{e}'_k\}$.

Note that *all the eigennumbers* λ_k, $k = 1, 2, \ldots, n$, *are nonzero.* Indeed, calculating $\det A$ for equation (7.99), we get $\det A = \lambda_1 \lambda_2 \ldots \lambda_n$ and since $\det A \neq 0$ for a central surface, it evidently follows that $\lambda_k \neq 0$.

* Recall that in passing from an orthonormal basis to an orthonormal one, the constant term c in the equation of the surface S does not change (see the third formula (7.87)).

Next we agree to enumerate all positive eigennumbers of the matrix A by the first index and all negative eigennumbers by the subsequent indices. Thus there is a number p such that

$$\lambda_1 > 0,\ \lambda_2 > 0,\ \ldots,\ \lambda_p > 0,$$
$$\lambda_{p+1} < 0,\ \lambda_{p+2} < 0,\ \ldots,\ \lambda_n < 0.$$

We now introduce the following designations:

if $\operatorname{sgn} \frac{\det B}{\det A} \neq 0$, we set

$$\left.\begin{aligned} \left|\frac{\det A}{\det B}\right| \lambda_k &= \frac{1}{a_k^2} \text{ for } k=1,\ 2,\ \ldots,\ p, \\ \left|\frac{\det A}{\det B}\right| \lambda_k &= -\frac{1}{a_k^2} \text{ for } k=p+1,\ \ldots,\ n, \end{aligned}\right\} \tag{7.100}$$

if $\operatorname{sgn} \frac{\det B}{\det A} = 0$, we set

$$\left.\begin{aligned} \lambda_k &= \frac{1}{a_k^2} \quad \text{for } k=1,\ 2,\ \ldots,\ p, \\ \lambda_k &= -\frac{1}{a_k^2} \quad \text{for } k=p+1,\ \ldots,\ n. \end{aligned}\right\} \tag{7.101}$$

Then equation (7.99) can evidently be rewritten as follows (we replace the designation of the coordinates x_k'' by x_k):

$$\frac{x_1^2}{a_1^2} + \frac{x_2^2}{a_2^2} + \ldots + \frac{x_p^2}{a_p^2} - \frac{x_{p+1}^2}{a_{p+1}^2} - \ldots - \frac{x_n^2}{a_n^2} + \operatorname{sgn} \frac{\det B}{\det A} = 0. \tag{7.102}$$

Equation (7.102) *is known as the* ***canonical equation*** *of the central second-order hypersurface.*

The quantities a_k, $k=1, 2, \ldots, n$, are called the *semi-axes of the central second-order hypersurface.* They can be calculated by formulas (7.100) and (7.101).

We use equation (7.102) to **classify** central hypersurfaces as follows.

1°. $p=n$, $\operatorname{sgn} \frac{\det B}{\det A} = -1$. In this case, the hypersurface S is called an **(n—1)-dimensional ellipsoid.**

The canonical equation of such an ellipsoid is usually written as

$$\frac{x_1^2}{a_1^2} + \ldots + \frac{x_n^2}{x_n^2} = 1. \tag{7.103}$$

If $a_1 = a_2 = \ldots = a_n = R$, then an $(n-1)$-dimensional ellipsoid is a sphere of radius R in an n-dimensional space.

Remark 1. In the case $p=0$, $\operatorname{sgn}\frac{\det B}{\det A}=1$, we also get an $(n-1)$-dimensional ellipsoid. It is evident that in this case equation (7.102) can be written in the form (7.103).

2°. $p=n$, $\operatorname{sgn}\frac{\det B}{\det A}=1$. The hypersurface is imaginary and is called an **imaginary ellipsoid.**

Remark 2. It is evident that in the case $p=0$, $\operatorname{sgn}\frac{\det B}{\det A}=-1$ we also obtain an imaginary ellipsoid.

3°. $0<p<n$, $\operatorname{sgn}\frac{\det B}{\det A}\neq 0$. In this case the central hypersurfaces are called **hyperboloids.**

The geometrical characteristics of a hyperboloid depend on the ratios of the numbers p and n and the value of $\operatorname{sgn}\frac{\det B}{\det A}$.

4°. $\operatorname{sgn}\frac{\det B}{\det A}=0$. In this case the central hypersurfaces are called **singular.** Among singular hypersurfaces, note the so-called **singular ellipsoid** corresponding to the values $p=0$ and $p=n$.

7.7.9. Simplification of the equation of a noncentral second-order hypersurface. Classification of noncentral hypersurfaces. Suppose that the hypersurface S, specified by equation (7.62) is not central, i.e.

$$\det A=0. \tag{7.104}$$

We carry out a standard simplification of equation (7.62). As a result, the equation assumes the form (7.98). We calculate $\det A$ using (7.98) (this is possible since $\det A$ is an invariant). Taking (7.104) into account, we obtain

$$\det A=\lambda_1\lambda_2\ldots\lambda_n=0.$$

Thus, at least one eigenvalue λ_k of the matrix A is zero. It should be emphasized that not all the eigenvalues are zero, since otherwise the quadratic form $A(\boldsymbol{x},\boldsymbol{x})$ would be identically equal to zero, whereas we assume (see 7.1.1) that the form is nonzero.

We leave in expression (7.98) only those terms in the first sum that correspond to the nonzero eigenvalues and then renumber the base vectors so that all the nonzero eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ correspond to the first p base vectors $\boldsymbol{e}'_1, \ldots, \boldsymbol{e}'_p$ (note that $p=\operatorname{rank} A$). After that, equation (7.98) can evidently be rewritten as

$$\sum_{k=1}^{p}\lambda_k x_k'^2+2\sum_{k=1}^{p}b'_k x'_k+2\sum_{k=p+1}^{n}b'_k x'_k+c=0 \tag{7.105}$$

(here $0<p<n$, $\lambda_1\neq 0, \ldots, \lambda_p\neq 0$; in addition, we especially isolated the first p terms of the second sum in equation (7.98)).

Let us now perform the following transformations.

1°. For every number k, $1\leqslant k\leqslant p$, we unite the terms with that number from the first and the second sum in (7.105) and then make the following transformations (taking into account that $\lambda_k\neq 0$):

$$\lambda_k x'^2_k+2b'_k x'_k=\lambda_k\left(x'^2_k+2\frac{b'_k}{\lambda_k}x'_k+\frac{b'^2_k}{\lambda_k^2}\right)$$
$$-\frac{b'^2_k}{\lambda_k}=\lambda_k\left(x'_k+\frac{b'_k}{\lambda_k}\right)^2-\frac{b'^2_k}{\lambda_k}.$$

After these transformations, (7.105) will evidently look like

$$\sum_{k=1}^{p}\lambda_k\left(x'_k+\frac{b'_k}{\lambda_k}\right)^2+2\sum_{k=p+1}^{n}b'_k x'_k+c'=0, \tag{7.106}$$

where the constant c' is defined by the equation

$$c'=c-\sum_{k=1}^{p}\frac{b'^2_k}{\lambda_k}. \tag{7.107}$$

Next we perform a translation by the formulas

$$x''_k=x'_k+\frac{b'_k}{\lambda_k},\ k=1,\ 2,\ \ldots,\ p;\ x''_k=x'_k,\ k=p+1,\ \ldots,\ n.$$

As a result, equation (7.106) passes into an equation

$$\sum_{k=1}^{p}\lambda_k x''^2_k+2\sum_{k=p+1}^{n}b'_k x''_k+c'=0, \tag{7.108}$$

where c' is defined by formula (7.107).

2°. We shall now seek a transformation of the orthonormal basis $\{e'_k\}$ under which the first p base vectors $e'_1, \ldots, e'_p$ do not change. At the expense of a change in the base vectors $e'_{p+1}, \ldots, e'_n$, we shall try to reduce the term $2\sum_{k=p+1}^{n} b'_k x''_k$ to the form $2\mu x'''_n$, where x'''_n is the nth coordinate in the new basis. Note that when we perform such a transformation, the constant term c' does not change.

It should be pointed out first that if all the coefficients b'_k in (7.108) are zero, then the aim of the transformation of 2° is attained, i.e. the term $2\sum_{k=p+1}^{n} b_k x''_k$ has the form $2\mu x''_n$, where $\mu=0$.

Thus we assume that at least one of the coefficients b'_k in the sum $\sum_{k=p+1}^{n} b'_k x''_k$ is nonzero. Then we can regard that sum as a certain linear form $B''(\boldsymbol{x})$ defined in the subspace V'' which is a linear closure of the vectors $\boldsymbol{e}'_{p+1}, \ldots, \boldsymbol{e}'_n$. According to the lemma of 5.4.1, this form can be represented in the indicated subspace as $B''(\boldsymbol{x}) = (\boldsymbol{h}, \boldsymbol{x})$, where $\boldsymbol{h}$ is some vector of the subspace V''. If now, in the subspace V'', we direct the unit vector $\boldsymbol{e}''_n$ along the vector $\boldsymbol{h}$ so that $\boldsymbol{h} = \mu \boldsymbol{e}''_n$, and choose the vectors $\boldsymbol{e}''_{p+1}, \ldots, \boldsymbol{e}''_{n-1}$ such that the system $\boldsymbol{e}''_{p+1}, \ldots, \boldsymbol{e}''_{n-1}, \boldsymbol{e}''_n$ is a basis in V'', then, evidently, $B''(\boldsymbol{x}) = (\boldsymbol{h}, \boldsymbol{x}) = \mu(\boldsymbol{e}''_n, \boldsymbol{x}) = \mu x'''_n$ in that basis since $(\boldsymbol{e}''_n, \boldsymbol{x}) = x'''_n$. Thus, choosing a basis in V'' in a way described above, we reduce $\sum_{k=p+1}^{n} b'_k x''_n$ to the form $\mu x'''_n$.

Thus, there is a transformation of the basis $\boldsymbol{e}'_1, \ldots, \boldsymbol{e}'_n$ into the orthonormal basis $\boldsymbol{e}''_1, \ldots, \boldsymbol{e}''_n$ (under which the vectors $\boldsymbol{e}'_1, \ldots, \boldsymbol{e}'_p$ remain unchanged) such that equation (7.108) assumes the form (we replace the designation of the coordinates x'''_k by x_k)

$$\sum_{k=1}^{p} \lambda_k x_k^2 + 2\mu x_n + c' = 0. \tag{7.109}$$

Note that the case $\mu = 0$ is not excluded in equation (7.109).

Equation (7.109) is called the ***canonical equation*** *of a noncentral second-order hypersurface.*

We can use equation (7.109) to **classify** noncentral hypersurfaces. Three cases are possible.

1°. $\mu \neq 0$, $p = \operatorname{rank} A = n - 1$.

In this case, we write the last two terms in equation (7.109) in the form $2\mu x_n + c' = 2\mu\left(x_n + \frac{c'}{2\mu}\right)$ and perform a translation along the axis x_n through the value $-c'/2\mu$. In order not to make the notation too cumbersome, we shall not change the designation of the coordinates in this case. As a result, the canonical equation (7.109) assumes the form

$$\lambda_1 x_1^2 + \ldots + \lambda_{n-1} x_{n-1}^2 + 2\mu x_n = 0. \tag{7.110}$$

The second-order hypersurfaces whose canonical equation has the form (7.110) are called **paraboloids.**

2°. $\mu = 0$, $p = \operatorname{rank} A < n$.

In this case, canonical equation (7.109) can be rewritten as

$$\lambda_1 x_1^2 + \ldots + \lambda_p x_p^2 + c' = 0. \tag{7.111}$$

In the subspace which is a linear closure of the vectors $e'''_1, \ldots, e'_p$ equation (7.111) is, evidently, the canonical equation of the central second-order surface S'. To get an idea of the hypersurface S in the whole space, we must place a plane, which is parallel to the plane V'' (a linear closure of the vectors $e'_{p+1}, \ldots, e'_n$) at each point of the surface S'. The locus of such planes forms a surface S. Thus the surface S is a **central cylinder** with a directing surface S' defined by equation (7.111) and generating surfaces which are parallel to the plane V''.

3°. $\mu \neq 0$, $p = \operatorname{rank} A < n-1$.

Acting as in 1°, we reduce the canonical equation (7.109) to the form

$$\lambda_1 x_1^2 + \ldots + \lambda_p x_p^2 + 2\mu x_n = 0. \tag{7.112}$$

In the subspace which is a linear closure of the vectors $e'_1, \ldots, e'_p, e'_n$, equation (7.112) evidently defines a paraboloid S' (see case 1°). To get an idea of the structure of the hypersurface S in the whole space, we must place a plane which is parallel to the plane V'' (a linear closure of the vectors $e'_{p+1}, \ldots, e'_{n-1}$) at each point of S'. The locus of such planes forms a surface S. Thus the surface S is a **paraboloidal cylinder** with a directing surface S' defined by equation (7.112) and generating planes which are parallel to the plane V''.

as the transformation of coordinates in transforming the orthonormal basis $e_1, e_2, \ldots, e_n, e_{n+1}$ of the $(n+1)$-dimensional Euclidean space, the matrix $\tilde{P}$ of that transformation having the form

$$\tilde{P}=\begin{pmatrix} p_{11} & \ldots & p_{n1} & 0 \\ & & & \cdot \\ p_{1n} & \ldots & p_{nn} & 0 \\ 0 & \ldots & 0 & 1 \end{pmatrix}. \tag{7.92}$$

It is easy to see that the matrix $\tilde{P}$ satisfies the condition $\tilde{P}'=\tilde{P}^{-1}$ and is, therefore, orthogonal. But then, according to 7.2.2, the orthonormal basis $e_1, e_2, \ldots, e_n, e_{n+1}$ is transformed into an orthonormal basis by means of the matrix $\tilde{P}$. We have found out above that under such transformation of the matrix B of the quadratic form, the determinant $\det B$ of that matrix is an invariant. We have proved the theorem.

Remark. It follows from the arguments presented in the proof of the theorem that the quantities rank A and rank B are also invariants of the general equation of a second-order hypersurface.

7.7.6. The centre of a second-order hypersurface. Let us try to find a translation under which general equation (7.76) would not contain the term $2B'(x')$ (or, if we turn to equation (7.82), the terms $2\sum_{k=1}^{n} b'_x x'_k$).

To put it otherwise, we shall seek a translation (i.e. the coordinates $\mathring{x}_1, \mathring{x}_2, \ldots, \mathring{x}_n$ of the point $\mathring{x}$) under which all the coefficients b'_k vanish. Directing our attention to formulas (7.84), we find that the required coordinates $\mathring{x}_1, \mathring{x}_2, \ldots, \mathring{x}_n$ of the point $\mathring{x}$ are a solution of the following system of linear equations:

$$\sum_{j=1}^{n} a_{jk}\mathring{x}_j+b_k=0, \quad k=1, 2, \ldots, n. \tag{7.93}$$

Equations (7.93) are known as the **equations of the centre of a second-order hypersurface,** and the point $\mathring{x}$ with the coordinates $(\mathring{x}_1, \mathring{x}_2, \ldots, \mathring{x}_n)$, where $(\mathring{x}_1, \mathring{x}_2, \ldots, \mathring{x}_n)$ is a solution of system (7.93), is called the **centre** of that surface.

Here is an explanation of the term the "centre" of a second-order hypersurface. Assume that the origin is at the centre $\mathring{x}$, i.e. the required translation has been performed. Then the equation of the surface S assumes the form

$$\sum_{j,\,k=1}^{n} a_{jk}x'_j x'_k+c'=0. \tag{7.94}$$

Proof. **(1) Necessity.** Suppose the vectors $e_1, e_2, \ldots, e_k$ are linearly dependent. Then one of them, say, e_k, is a linear combination of the other vectors:

$$e_k = \alpha_1 e_1 + \alpha_2 e_2 + \ldots + \alpha_{k-1} e_{k-1}.$$

Performing a scalar multiplication of the last relation by e_i, $i=1, 2, \ldots, k$, we find that the last row of Gramian (8.1) is a linear combination of the first $k-1$ rows. By Theorem 1.7, this determinant is zero. We have proved the necessity.

(2) Sufficiency. We assume that Gramian (8.1) is zero. Then its columns are linearly dependent, i.e. there are numbers $\beta_1, \beta_2, \ldots, \beta_k$ which are not all equal to zero and such that the relations

$$\beta_1 (e_i, e_1) + \beta_2 (e_1, e_2) + \ldots + \beta_k (e_i, e_k) = 0$$

are satisfied for $i=1, 2, \ldots, k$. Rewriting these relations in the form

$$(e_i, \beta_1 e_1 + \beta_2 e_2 + \ldots + \beta_k e_k) = 0, \quad i=1, 2, \ldots, k,$$

we make sure that the vector $\beta_1 e_1 + \beta_2 e_2 + \ldots + \beta_k e_k$ is orthogonal to all the vectors $e_1, e_2, \ldots, e_k$, i.e. is orthogonal to the linear closure L of those vectors. Since this vector belongs to L, it is equal to zero. Since not all the β_j are zero, this means that the vectors $e_1, e_2, \ldots, e_k$ are linearly dependent. And this is what we set out to prove.

Corollary. *If the vectors $e_1, e_2, \ldots, e_k$ are linearly independent, then the Gramian of those vectors is nonzero.*

Let us prove that in this case the Gramian is **positive.** Let L be a linear closure of the vectors $e_1, e_2, \ldots, e_k$. It is evident that $e_1, e_2, \ldots, e_k$ is a basis in L. Let us consider a bilinear symmetric form $A(x, y)$ which is a scalar product (x, y): $A(x, y) = (x, y)$. The corresponding quadratic form $A(x, x) = (x, x)$ is evidently a definite form and, therefore, in accordance with Theorem 7.6 (Sylvester's criterion), the determinant $\det(a_{ij})$ of its matrix (a_{ij}) in the basis $e_1, e_2, \ldots, e_k$ is positive. But this determinant is precisely Gramian (8.1) of the system $e_1, e_2, \ldots, e_k$ since $a_{ij} = (e_i, e_j)$.

8.1.2. Reciprocal bases. Covariant and contravariant vector coordinates. Let $e_1, e_2, \ldots, e_n$ be a basis in the Euclidean space E^n. The basis $e^1, e^2, \ldots, e^n$ is said to be a **reciprocal** of the basis e_i, $i=1, 2, \ldots, n$, if the relations

$$(e_i, e^j) = \delta_i^j = \begin{cases} 1 & \text{for } i=j \\ 0 & \text{for } i \neq j \end{cases} \tag{8.2}$$

are satisfied for $i, j=1, 2, \ldots, n$.

The symbol δ_i^j is known as the **Kronecker delta.**

A question arises whether a reciprocal basis exists and is unique. The question is answered in the affirmative: *there is a unique reciprocal basis for any given basis $e_1, e_2, \ldots, e_n$.* To prove this statement,

we do the following. We assume that $x_1^j, x_2^j, \ldots, x_n^j$ are the coordinates of the required vectors $\boldsymbol{e}^j$ in the basis $\boldsymbol{e}_i$:

$$\boldsymbol{e}^j = x_1^j \boldsymbol{e}_1 + x_2^j \boldsymbol{e}_2 + \ldots + x_n^j \boldsymbol{e}_n, \quad j = 1, 2, \ldots, n. \tag{8.3}$$

Multiplying scalarly both sides of the last relations by $\boldsymbol{e}_i$, we get, using (8.2),

$$x_1^j (\boldsymbol{e}_i, \boldsymbol{e}_1) + x_2^j (\boldsymbol{e}_1, \boldsymbol{e}_2) + \ldots + x_n^j (\boldsymbol{e}_i, \boldsymbol{e}_n) = \delta_i^j \tag{8.4}$$
$$i, j = 1, 2, \ldots, n.$$

For a fixed j, relations (8.4) can be regarded as a quadratic system of linear equations with respect to the unknown coordinates x_1^j, $x_2^j, \ldots, x_n^j$ of the vector $\boldsymbol{e}^j$ in the basis $\boldsymbol{e}_i$. Since the determinant of system (8.4) is the Gramian of the base vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$, it follows, in accordance with the corollary of Theorem 8.1, that it is nonzero and, therefore, system (8.4) has a unique solution $x_1^j, x_2^j, \ldots, x_n^j$ which is nonzero, since the system is nonhomogeneous. Then we construct vectors $\boldsymbol{e}^j$ with the aid of relations (8.3) which evidently satisfy relations (8.2).

It remains to verify that the vectors $\boldsymbol{e}^1, \boldsymbol{e}^2, \ldots, \boldsymbol{e}^n$ form a basis.

Suppose a certain linear combination of these vectors is equal to zero: $\alpha_1 \boldsymbol{e}^1 + \alpha_2 \boldsymbol{e}^2 + \ldots + \alpha_n \boldsymbol{e}^n = 0$.

Multiplying scalarly the last equation in succession by $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots$ $\ldots, \boldsymbol{e}_n$ and using equations (8.2), we get $\alpha_1 = 0, \alpha_2 = 0, \ldots, \alpha_n = 0$. Consequently, the vectors $\boldsymbol{e}^1, \boldsymbol{e}^2, \ldots, \boldsymbol{e}^n$ are linearly independent, i.e. form a basis.

Thus the reciprocal basis $\boldsymbol{e}^j$ of the basis $\boldsymbol{e}_i$ exists and is defined uniquely.

Remark 1. By virtue of the symmetry of relations (8.2) with respect to $\boldsymbol{e}_i$ and $\boldsymbol{e}^j$ the basis $\boldsymbol{e}_i$ is a reciprocal basis of $\boldsymbol{e}^j$. Therefore, in what follows we shall speak of the reciprocal bases $\boldsymbol{e}_i$ and $\boldsymbol{e}^j$.

Remark 2. If the basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ is orthonormal, then the reciprocal basis $\boldsymbol{e}^j$ coincides with the given basis. Indeed, setting $\boldsymbol{e}^j = \boldsymbol{e}_j$ in this case, we make sure that relations (8.2) are satisfied. Using the property of uniqueness of the reciprocal basis, we verify the validity of the remark.

Assume that $\boldsymbol{e}_i$, $\boldsymbol{e}^j$ are reciprocal bases and $\boldsymbol{x}$ is an arbitrary vector of the space. Resolving the vector $\boldsymbol{x}$ into the components $\boldsymbol{e}_i$ and $\boldsymbol{e}^j$, we get

$$\left.\begin{aligned} \boldsymbol{x} &= x_1 \boldsymbol{e}^1 + x_2 \boldsymbol{e}^2 + \ldots + x_n \boldsymbol{e}^n, \\ \boldsymbol{x} &= x^1 \boldsymbol{e}_1 + x^2 \boldsymbol{e}_2 + \ldots + x^n \boldsymbol{e}_n. \end{aligned}\right\} \tag{8.5}$$

The coordinates $(x_1, x_2, \ldots, x_n)$ of the vector $\boldsymbol{x}$ in the basis $\boldsymbol{e}^j$ are said to be **covariant coordinates** of the vector $\boldsymbol{x}$, and the coordinates $(x^1, x^2, \ldots, x^n)$ of that vector in the basis $\boldsymbol{e}_i$ are said to be **contra-**

variant coordinates of the vector $\boldsymbol{x}$. We shall explain these terminology in 8.1.3.

To abbreviate the notation of the formulas which contain summands of the same type (relations (8.5) can serve as an example of such formulas), we shall use the agreement on summation in what follows. The agreement consists in the following. Suppose we have an expression formed from factors with a finite number of indices, some of which are subscripts and the others are superscripts. We agree to denote all the subscripts by different symbols and all the superscripts also by different symbols. If there are two similar indices in the expression, one of which is a superscript and the other is a subscript, then it is considered that a summation is performed by these indices, i.e. the indices are successively enumerated as 1, 2, ..., n and the results are summed up.

For example,

$$x_i\boldsymbol{e}^i = x_1\boldsymbol{e}^1 + x_2\boldsymbol{e}^2 + \ldots + x_n\boldsymbol{e}^n, \quad \delta_i^i = \delta_1^1 + \delta_2^2 + \ldots + \delta_n^n,$$

$$\begin{aligned} g_{ij}x^ix^j &= (g_{1j}x^1x^j) + (g_{2j}x^2x^j) + \ldots + (g_{nj}x^nx^j) \\ &= (g_{11}x^1x^1 + g_{12}x^1x^2 + \ldots + g_{1n}x^1x^n) \\ &\quad + (g_{21}x^2x^1 + g_{22}x^2x^2 + \ldots + g_{2n}x^2x^n) \\ &\quad \ldots\ldots\ldots\ldots\ldots\ldots \\ &\quad + (g_{n1}x^nx^1 + g_{n2}x^nx^2 + \ldots + g_{nn}x^nx^n). \end{aligned}$$

The agreement on summation makes it possible to write formulas (8.5) in the following compact form:

$$\boldsymbol{x} = x_i\boldsymbol{e}^i, \quad \boldsymbol{x} = x^i\boldsymbol{e}_i. \tag{8.6}$$

Remark 3. Similar superscripts and subscripts which were mentioned in the agreement on summation are usually called the *sum indices*. It is clear that we can denote the sum indices by any identical symbols without changing the expression in which they appear. For instance, $x_i\boldsymbol{e}^i$ and $x_j\boldsymbol{e}^j$ constitute the same expression.

Let us obtain an explicit expression for the covariant and contravariant coordinates of the vector $\boldsymbol{x}$.

For that purpose, we multiply scalarly the first equation (8.6) by $\boldsymbol{e}_j$ and the second equation by $\boldsymbol{e}^j$. Taking into account relations (8.2), we find that

$$(\boldsymbol{x}, \boldsymbol{e}_j) = x_i(\boldsymbol{e}^i, \boldsymbol{e}_j) = x_i\delta_j^i = x_j, \quad (\boldsymbol{x}, \boldsymbol{e}^j) = x^i(\boldsymbol{e}_i, \boldsymbol{e}^j) = x^i\delta_i^j = x^j.$$

Thus we have

$$x_j = (\boldsymbol{x}, \boldsymbol{e}_j), \quad x^j = (\boldsymbol{x}, \boldsymbol{e}^j). \tag{8.7}$$

Using relations (8.7), we write formulas (8.6) as

$$\boldsymbol{x} = (\boldsymbol{x}, \boldsymbol{e}_i)\boldsymbol{e}^i, \quad \boldsymbol{x} = (\boldsymbol{x}, \boldsymbol{e}^i)\boldsymbol{e}_i. \tag{8.8}$$

Relations (8.8) are known as the **Gibbs* formulas.**

Let us again direct our attention to the problem of constructing reciprocal bases.

With the aid of formulas (8.8) we get

$$\boldsymbol{e}^i=(\boldsymbol{e}^i,\ \boldsymbol{e}^j)\boldsymbol{e}_j,\ \boldsymbol{e}_i=(\boldsymbol{e}_i,\ \boldsymbol{e}_j)\boldsymbol{e}^j. \tag{8.9}$$

We introduce the designations

$$g_{ij}=(\boldsymbol{e}_i,\ \boldsymbol{e}_j),\ g^{ij}=(\boldsymbol{e}^i,\ \boldsymbol{e}^j). \tag{8.10}$$

Using these designations, we rewrite relations (8.9) as

$$\boldsymbol{e}^i=g^{ij}\boldsymbol{e}_j,\ \boldsymbol{e}_i=g_{ij}\boldsymbol{e}^j. \tag{8.11}$$

Thus, *to construct the basis* $\boldsymbol{e}^i$ *from the basis* $\boldsymbol{e}_i$, *it is sufficient to know the matrix* (g^{ij}), *and to construct the basis* $\boldsymbol{e}_i$ *from the basis* $\boldsymbol{e}^i$, *it is sufficient to know the matrix* (g_{ij}).

Let us prove that the indicated matrices are mutually inverse. Note that since the elements of the inverse matrix can be calculated in terms of the elements of the given matrix, it is clear that the problem of constructing reciprocal bases can be solved with the aid of relations (8.11).

Let us ascertain that the matrices (g^{ij}) and (g_{ij}) are mutually inverse.

Multiplying scalarly the first equation (8.11) by $\boldsymbol{e}_k$, we get $(\boldsymbol{e}^i, \boldsymbol{e}_k)=g^{ij}(\boldsymbol{e}_j,\ \boldsymbol{e}_k)$. Taking (8.2) and (8.10) into account, we find from this relation that

$$g^{ij}g_{jk}=\delta^i_k=\begin{cases}1 & \text{for } i=k,\\ 0 & \text{for } i\neq k.\end{cases}$$

Thus, the product of the matrices (g^{ij}) and (g_{ij}) is an identity matrix. Consequently, the matrices (g^{ij}) and (g_{ij}) are mutually inverse.

8.1.3. Transformation of a basis and coordinates. Suppose $\boldsymbol{e}_i$ and $\boldsymbol{e}^i$ are given reciprocal bases and $\boldsymbol{e}_i$ and $\boldsymbol{e}^{i'}$ are some new reciprocal bases whose elements are denoted by primed indices. In fact this means that we introduce a new natural series $1', 2', 3', \ldots$ and suppose that the index i' assumes the values $1', 2', \ldots, n'$. Thus the indices i and i' independently assume different values

$$i=1,\ 2,\ \ldots,\ n,\ i'=1',\ 2',\ \ldots,\ n'.$$

Using the agreement on summation introduced in 8.1.2, we write the formulas for the transformation of base vectors. As a result we get:

(1) formulas for transition from the old basis $\boldsymbol{e}_i$ to a new basis $\boldsymbol{e}^{i'}$ and formulas for a reverse transition

* Gibbs, Josian Willard (1839-1903), an American mathematical physicist.

$$e_{i'}=b^i_{i'}e_i,\ e_{i'}=b^{i'}_i e_{i'},\ i=1,\ 2,\ \ldots,\ n,\ i'=1',\ 2',\ \ldots,\ n', \quad (8.12)$$

(2) formulas for transition from the old basis e^i to a new basis $e^{i'}$ and formulas for a reverse transition:

$$e^{i'}=\tilde{b}^{i'}_i e^i,\ e^i=\tilde{b}^i_{i'}e^{i'},\ i=1,\ 2,\ \ldots,\ n,\ i'=1',\ 2',\ \ldots,\ n'. \quad (8.13)$$

Since transformations (8.12) (as well as (8.13)) are mutually reciprocal, the matrices $(b^i_{i'})$ and $(b^{i'}_i)$ (as well as $(\tilde{b}^i_{i'})$ and $(\tilde{b}^{i'}_i)$) are mutually inverse.

Let us prove that the matrices $(b^i_{i'})$ and $(\tilde{b}^i_{i'})$ are identical. We shall thus prove that the matrices $(b^{i'}_i)$ and $(\tilde{b}^{i'}_i)$ are identical. To prove this, we multiply scalarly the first equation (8.12) by e^k and the second equation (8.13) by $e_{k'}$. Taking relations (8.2) into account, we obtain

$$(e_{i'},\ e^k)=b^i_{i'}(e_i,\ e^k)=b^i_{i'}\delta^k_i=b^k_{i'},$$
$$(e^i,\ e_{k'})=\tilde{b}^i_{i'}(e^{i'},\ e_{k'})=\tilde{b}^i_{i'}\delta^{i'}_{k'}=\tilde{b}^i_{k'}.$$

From these relations we get for $k=i$, $k'=i'$ the following equations:

$$b^i_{i'}=(e_{i'},\ e^i), \quad (8.14)$$
$$\tilde{b}^i_{i'}=(e_{i'},\ e^i). \quad (8.15)$$

Since the right-hand sides of relations (8.14) and (8.15) are equal, the left-hand sides are equal too. In other words, $b^i_{i'}=\tilde{b}^i_{i'}$ and this means that the matrices $(b^i_{i'})$ and $(\tilde{b}^i_{i'})$ are identical. Note that the elements $b^i_{i'}$ of the matrix $(b^i_{i'})$ can be calculated by formulas (8.14).

Thus, the following statement holds true.

To pass from the basis e_i, e^i to the basis $e_{i'}$, $e^{i'}$, it is sufficient to know the matrix $(b^i_{i'})$ of transition from the basis e_i to the basis $e_{i'}$ (the matrix $(b^{i'}_i)$ can be calculated from the matrix $(b^i_{i'})$).

Here is a full collection of formulas for the transformation of base vectors:

$$\left.\begin{aligned} e_{i'}&=b^i_{i'}e_i,\ e_i=b^{i'}_i e_{i'},\\ e^{i'}&=b^{i'}_i e^i,\ e^i=b^i_{i'}e^{i'}. \end{aligned}\right\} \quad (8.16)$$

Let us derive formulas for the transformation of the coordinates of the vector x in passing to a new basis.

Let $x_{i'}$ be the covariant coordinates of the vector x in the basis $e_{i'}$, $e^{i'}$. Then, according to (8.7), we have $x_{i'}=(x,\ e_{i'})$. Substituting the expression for $e_{i'}$ from formulas (8.16) into the right-hand side of this relation, we find that $x_{i'}=(x,\ b^i_{i'}e_i)=b^i_{i'}(x,\ e_i)=b^i_{i'}x_i$.

We arrive at the following conclusion: *when we pass to a new ba-*

sis, the formulas for the transformation of the covariant coordinates of the vector $\boldsymbol{x}$ *have the form*

$$x_{i'} = b^{i}_{i'} x_i. \tag{8.17}$$

Consequently, *when we pass to a new basis, the covariant coordinates of the vector* $\boldsymbol{x}$ *are transformed by means of the matrix* $(b^{i}_{i'})$ *of a direct transition from the old basis to a new one.*

This coherence of transformations explains the name of the *covariant** coordinates of a vector.

Let us now consider the transformation of the contravariant coordinates of the vector $\boldsymbol{x}$.

Substituting the expression for $\boldsymbol{e}^{i'}$ from formulas (8.16) into the right-hand side of the relation $x^{i'} = (\boldsymbol{x}, \boldsymbol{e}^{i'})$, we find, after some transformations, that

$$x^{i'} = b^{i'}_{i} x^i. \tag{8.18}$$

We see that *when we pass to a new basis, the contravariant coordinates of the vector* $\boldsymbol{x}$ *are transformed by means of the matrix* $(b^{i'}_{i})$ *of the reverse transition from the new basis to the old one.*

This nonconcordance of the transformations explains the name of the *contravariant*** coordinates of a vector.

8.2. The Concept of a Tensor. Principal Operations on Tensors

8.2.1. The concept of a tensor. We discuss here an arbitrary (not necessarily Euclidean) real n-dimensional linear space L^n.

Definition. *A* ***tensor*** *A of the type* (p, q) *(p times covariant and q times contravariant) is a geometrical object which* (1) *is defined by the* n^{p+q} *coordinates* $A^{k_1 \ldots k_q}_{i_1 \ldots i_p}$ *in every basis* $\boldsymbol{e}_i$ *of the linear space* L^n *(the indices* $i_1, \ldots, i_p, k_1, \ldots, k_q$ *independently assume the values* $1, 2, \ldots, n$*),* (2)*, possesses the property that its coordinates* $A^{k'_1 \ldots k'_q}_{i'_1 \ldots i'_p}$ *in the basis* $\boldsymbol{e}_{i'}$ *are connected with the coordinates* $A^{k_1 \ldots k_q}_{i_1 \ldots i_p}$ *in the basis* $\boldsymbol{e}_i$ *by the relations*

$$A^{k'_1 \ldots k'_q}_{i'_1 \ldots i'_p} = b^{i_1}_{i'_1} \ldots b^{i_p}_{i'_p} b^{k'_1}_{k_1} \ldots b^{k'_q}_{k_q} A^{k_1 \ldots k_q}_{i_1 \ldots i_p} \ . \tag{8.19}$$

in which $b^{i}_{i'}$ *are the elements of the matrix* $(b^{i}_{i'})$ *of the transition from the basis* $\boldsymbol{e}_i$ *to the basis* $\boldsymbol{e}_{i'}$ *and* $b^{k'}_{k}$ *are the elements of the matrix of the reverse transition from* $\boldsymbol{e}_{i'}$ *to* $\boldsymbol{e}_i$.

---

* Covariant means concordantly changing.

** Contravariant means oppositely changing.

The number $r=p+q$ is called the ***rank*** *of the tensor.*

Remark 1. Formulas (8.19) are called the formulas for the transformation of the coordinates of a tensor in the transformation of a basis.

Note that the covariant and contravariant coordinates of a vector aretransformed by formulas (8.19) (for $p=1$ and $q=0$ in the first case and for $p=0$ and $q=1$ in the second case, see 8.1.3.). Therefore, *a vector is a tensor of rank* 1 (once covariant or oncecontravariant depending on the choice of the type of coordinates of the vector).

Note that there are also *tensors of rank* 0. These are tensors having only one coordinate, the coordinate haviqg no indices and being of the same value in all the coordinate systems. The tensors of rank 0 are usually called **invariant tensors.**

Remark 2. The indices $i_1, \ldots, i_p$ are said to be covariant and $k_1 \ldots, k_q$, contravariant. The name can be explained by the fact that with respect of each index the transformation of the coordinates of a tensor is carried out by a complete analogy with the transformations of the covariant and contravariant coordinates of a vector (see formulas (8.17) and (8.18)).

To make the definition of a tensor correct, we must make sure that the successive transitions from the basis $\boldsymbol{e}_i$ to the basis $\boldsymbol{e}_{i'}$ and then m from $\boldsymbol{e}_{i'}$ to $\boldsymbol{e}_{i''}$ can be reduced to the same transformation of the coordinates of a tensor as in the direct transition from $\boldsymbol{e}_i$ to $\boldsymbol{e}_{i''}$.

Let $(b^i_{i'})$, $(b^{i'}_{i''})$ and $(b^i_{i''})$ be, respectively, the matrices of the transition from the basis $\boldsymbol{e}_i$ to the basis $\boldsymbol{e}_{i'}$, from $\boldsymbol{e}_{i'}$ to $\boldsymbol{e}_{i''}$ and from $\boldsymbol{e}_i$ to $\boldsymbol{e}_{i''}$. Since the transformation matrices are multiplied in successive transitions, the following relations are obvious:

$$b^i_{i''}=b^{i'}_{i''}b^i_{i'}, \quad b^{i''}_i=b^{i''}_{i'}b^{i'}_i. \tag{8.20}$$

After the remarks we have made, let us verify the correctness of the definition of a tensor. Let $A^{k_1\ldots k_q}_{i_1\ldots i_p}$, $A^{k'_1\ldots k'_q}_{i'_1\ldots i'_p}$, $A^{k''_1\ldots k''_q}_{i''_1\ldots j''_p}$ be the coordinates of the tensor A in the basis $\boldsymbol{e}_i$, $\boldsymbol{e}_{i'}$ and $\boldsymbol{e}_{i''}$ respectively. By formulas (8.19) for the successive transition from $\boldsymbol{e}_i$ to $\boldsymbol{e}_{i'}$ and then to $\boldsymbol{e}_{i''}$ we get

$$A^{k'_1\ldots k'_q}_{i'_1\ldots i'_p}=b^{i_1}_{i'_1}\ldots b^{i_p}_{i'_p}b^{k'_1}_{k_1}\ldots b^{k'_q}_{k_q}A^{k_1\ldots k_q}_{i_1\ldots i_p}, \tag{8.21}$$

$$A^{k''_1\ldots k''_q}_{i''_1\ldots i''_p}=b^{i'_1}_{i''_1}\ldots b^{i'_p}_{i''_p}b^{k''_1}_{k'_1}\ldots b^{k''_q}_{k'_q}A^{k'_1\ldots k'_q}_{i'_1\ldots i'_p}. \tag{8.22}$$

Substituting the expression $A^{k'_1\ldots k'_q}_{i'_1\ldots i'_p}$ for the coordinates from

(8.21) and taking relations (8.20) into account, we obtain

$$A^{k''_1 \dots k''_q}_{i''_1 \dots i''_p} = \left(b^{i'_1}_{i''_1} b^{i_1}_{i'_1}\right) \dots \left(b^{i'_p}_{i''_p} b^{i_p}_{i'_p}\right) \left(b^{k''_1}_{k'_1} b^{k'_1}_{k_1}\right)$$

$$\dots \left(b^{k''_q}_{k'_q} b^{k'_q}_{k_q}\right) A^{k_1 \dots k_q}_{i_1 \dots i_p} = b^{i_1}_{i''_1} \dots b^{i_p}_{i''_p} b^{k''_1}_{k_1} \dots b^{k''_q}_{k_q} A^{k_1 \dots k_q}_{i_1 \dots i_{q_p}}$$

Thus consecutive transitions from the basis $\boldsymbol{e}_i$ to the basis $\boldsymbol{e}_{i'}$ and then to the basis $\boldsymbol{e}_{i''}$ lead to the same transformation of the coordinates of a tensor as in a direct transition from $\boldsymbol{e}_i$ to $\boldsymbol{e}_{i''}$. We have established the correctness of the definition of a tensor.

Remark 3. Any system n^{p+q} of the numbers $A^{k_1 \dots k_q}_{i_1 \dots i_p}$ can be regarded in the given basis $\boldsymbol{e}_i$ as the coordinates of a certain tensor A of type (p, q). To verify this fact, we define in an arbitrary basis $\boldsymbol{e}_{i'}$ by means of formulas (8.19), a system of numbers $A^{k'_1 \dots k'_q}_{i'_1 \dots i'_p}$ which will be regarded as the coordinates of the required tensor A in the basis $\boldsymbol{e}_{i'}$. These coordinates are evidently transformed by formulas (8.19) when we pass from the basis $\boldsymbol{e}_i$ to the basis $\boldsymbol{e}_{i'}$. As before, it is easy to verify that the consecutive transitions from the basis $\boldsymbol{e}_i$ to the basis $\boldsymbol{e}_{i'}$ and then to the basis $\boldsymbol{e}_{i''}$ lead to the same transformation of the coordinates obtained as in a direct transition from $\boldsymbol{e}_i$ to $\boldsymbol{e}_{i''}$. Consequently, the system of numbers $A^{k_1 \dots k_q}_{i_1 \dots i_p}$ indeed constitutes the coordinates of a certain tensor A of type (p, q).

8.2.2. Examples of tensors.

1°. **A null-tensor.** Among the tensors of type (p, q) we must isolate the so-called *null-tensor*. This is a tensor whose coordinates in any basis are zero. Relations (8.19) are evidently satisfied.

Note that if the coordinates of the tensor A are zero in some basis, then, according to (8.19), they are zero in any basis and, consequently, A is a null-tensor.

2°. **The Kronecker delta.** Let us verify that the tensor A of type $(1, 1)$, having the coordinates δ^k_i iin the basis $\boldsymbol{e}_i$, has the coordinates $\delta^{k'}_{i'}$ in the basis $\boldsymbol{e}_{i'}$.

Thus, let A be a tensor having the coordinates δ^k_i in the basis $\boldsymbol{e}_i$. To find the coordinates of this tensor in the basis $\boldsymbol{e}_{i'}$, we must make use of formulas (8.19), i.e. the coordinates of the tensor A in the basis $\boldsymbol{e}_{i'}$, are equal to $b^{k'}_k b^i_{i'} \delta^k_i$. Using the properties of the Kronecker delta, we find that $b^{k'}_k b^i_{i'} \delta^k_i = b^{k'}_k b^k_{i'} = \delta^{k'}_{i'}$.

Thus, in the new basis $\boldsymbol{e}_{i'}$ the coordinates of the tensor A are indeed equal to $\delta^{k'}_{i'}$. We can, therefore, regard the Kronecker delta as a tensor of type $(1, 1)$.

3°. Suppose $A(\boldsymbol{x}, \boldsymbol{y})$ is a bilinear form defined in a finite-dimensional Euclidean space E^n, and $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ is some basis in that space. Then the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ can be represented in the form $\boldsymbol{x}=x^i\boldsymbol{e}_i$, $\boldsymbol{y}=y^j\boldsymbol{e}_j$.

Using the linear property of the form $A(\boldsymbol{x}, \boldsymbol{y})$ with respect to each argument, we can write

$$A(\boldsymbol{x}, \boldsymbol{y})=A(x^i\boldsymbol{e}_i, y^j\boldsymbol{e}_j)=A(\boldsymbol{e}_i, \boldsymbol{e}_j)x^iy^j.$$

Designating $A(\boldsymbol{e}_i, \boldsymbol{e}_j)$ as a_{ij}, we get

$$a_{ij}=A(\boldsymbol{e}_i, \boldsymbol{e}_j). \tag{8.23}$$

Then the form $A(\boldsymbol{x}, \boldsymbol{y})$ can be written as follows:

$$A(\boldsymbol{x}, \boldsymbol{y})=a_{ij}x^iy^j. \tag{8.24}$$

Let us ascertain that the coefficients a_{ij} of the matrix having the form $A(\boldsymbol{x}, \boldsymbol{y})$ are transformed by the law (8.19) of transformation of the coordinates of a tensor of type (2, 0) when we pass to a new basis, i.e. they constitute a tensor of type (2, 0).

Let us consider an arbitrary basis $\boldsymbol{e}_{1'}, \boldsymbol{e}_{2'}, \ldots, \boldsymbol{e}_{n'}$. We write the form $A(\boldsymbol{x}, \boldsymbol{y})$ in this basis in the form (8.24), i.e. $A(\boldsymbol{x}, \boldsymbol{y})=a_{i'j'}x^{i'}y^{j'}$, with

$$a_{i'j'}=A(\boldsymbol{e}_{i'}, \boldsymbol{e}_{j'}). \tag{8.25}$$

We pass from the basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ to a new basis $\boldsymbol{e}_{1'}, \boldsymbol{e}_{2'}, \ldots, \boldsymbol{e}_n$. Designating the transition matrix from the basis $\boldsymbol{e}_i$ to the basis $\boldsymbol{e}_{i'}$ as $b^i_{i'}$, we get $\boldsymbol{e}_{i'}=b^i_{i'}\boldsymbol{e}_i$, $\boldsymbol{e}_{j'}=b^j_{j'}\boldsymbol{e}_j$.

Substituting these expressions for $\boldsymbol{e}_{i'}$ and $\boldsymbol{e}_{j'}$ into the right-hand side of (8.25) and using the linear property of the form $A(\boldsymbol{x}, \boldsymbol{y})$ with respect to each argument, we find that $a_{i'j'}=A(b^i_{i'}\boldsymbol{e}_i, b^j_{j'}\boldsymbol{e}_j)=b^i_{i'}b^j_{j'}A(\boldsymbol{e}_i, \boldsymbol{e}_j)$.

In accordance with formula (8.23), the last relation can be rewritten as $a_{i'j'}=b^i_{i'}b^j_{j'}a_{ij}$.

Consequently, the coefficients a_{ij} of the matrix of a bilinear form are transformed by the law (8.19) of transformation of the coordinates of a tensor of type (2,0) and can, therefore, be regarded as the coordinates of a tensor of that type.

4°. Every linear operator given in a finite-dimensional Euclidean space E^n and acting into the same space can be put into correspondence with a certain tensor of type (1, 1), and this tensor completely defines the indicated operator.

Suppose $\boldsymbol{y}=L\boldsymbol{x}$ is a linear operator defined in E^n, and $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ is a basis in E^n. Since $\boldsymbol{x}=x^i\boldsymbol{e}_i$ and $\boldsymbol{y}=y^j\boldsymbol{e}_j$ and L is a linear operator, we have

$$y^j\boldsymbol{e}_j=x^jL(\boldsymbol{e}_i). \tag{8.26}$$

We resolve the vector $L(\boldsymbol{e}_i)$ into the components $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$.

$$L(\boldsymbol{e}_i)=a_i^j\boldsymbol{e}_j.$$

Substituting the expression obtained for $L(\boldsymbol{e}_i)$ into (8.26) and using the uniqueness of resolution into components, we get

$$y^j=a_i^jx^i,\ j=1,\ 2,\ \ldots,\ n. \tag{8.27}$$

Recall that relations (8.27) can be regarded as the coordinate method of defining a linear operator. Then the matrix (a_i^j) of the coefficients a_i^j is called the matrix of a linear operator.

Let us ascertain that the coefficients of this matrix are transformed by the law (8.19) of transformation of the coordinates of a tensor of type (1, 1) when we pass to a new basis and, therefore, constitute a tensor of type (1, 1).

Let us consider an arbitrary basis $\boldsymbol{e}_{1'}, \boldsymbol{e}_{2'}, \ldots, \boldsymbol{e}_{n'}$. We write the linear operator L in this basis in the form (8.27):

$$y^{j'}=a_{i'}^{j'}x^{i'},\ j'=1',\ 2',\ \ldots,\ n'. \tag{8.28}$$

We pass from the basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ to the basis $\boldsymbol{e}_{1'}, \boldsymbol{e}_{2'}, \ldots, \boldsymbol{e}_{n'}$. Designating the transition matrix $(b_{i'}^i)$ (or, what is the same, $(b_{j'}^j)$), we get* (see 8.1.3)

$$x^i=b_{i'}^ix^{i'},\ y^j=b_{k'}^jy^{k'}.$$

We substitute these expressions for x^i and y^j into (8.27) and get the following relations:

$$y^{k'}b_{k'}^j=a_i^jb_{i'}^ix^{i'},\ j=1,\ 2,\ \ldots,\ n. \tag{8.29}$$

We must get expression for $y^{j'}$ from (8.29). For that purpose, we multiply both sides of (8.29) by $b_j^{j'}$ and perform a summation by j from 1 to n. Taking into consideration that $b_{k'}^jb_j^{j'}=\delta_{k'}^{j'}$, we get $y^{k'}\delta_{k'}^{j'}=(b_{i'}^ib_j^{j'}a_i^j)\,x^{i'}$. Note that $y^{k'}\ \delta_{k'}^{j'}=y^{j'}$. Therefore $y^{j'}=(b_{i'}^ib_j^{j'}a_i^j)x^{i'}$. Comparing this expression for $y^{j'}$ with the expression for $y^{j'}$ obtained by formula (8.28), we get the following identity which is valid for any vectors $\boldsymbol{x}$ (for any coordinates $x^{i'}$):

$$a_{i'}^{j'}x^{i'}=(b_{i'}^ib_j^{j'}a_i^j)\,x^{i'}.$$

It follows from this and from the arbitrariness of $x^{i'}$ that the coefficients a_i^j of the matrix of a linear operator are transformed by the law $a_{i'}^{j'}=b_{i'}^ib_j^{j'}a_i^j$.

Thus, the coefficients a_i^j are transformed by the law (8.19) of transformation of the coordinates of a tensor of type (1, 1) and, therefore, constitute such a tensor.

8.2.3. The principal operations on tensors. The principal operations on tensors are the *addition and subtraction of tensors, multi-*

---

* We designate the sum index in the formula for y^j as k'.

plication of a tensor by a number, multiplication of tensors, contraction of tensors, commutation of indices, symmetrization and alternation of tensors.

Let us define these operations.

1°. **Addition and subtraction of tensors.** The operations of addition and subtraction are defined for tensors of the same type.

Suppose A and B are two tensors of the same type (p, q), and $A^{k_1 \ldots k_q}_{i_1 \ldots i_p}$ and $B^{k_1 \ldots k_q}_{i_1 \ldots i_p}$ are the coordinates with like indices of these tensors in the basis $\boldsymbol{e}_i$.

The **sum** $(A+B)$ (the **difference** $A-B$) of these tensors is a tensor having the coordinates

$$A^{k_1 \ldots k_q}_{i_1 \ldots i_p} + B^{k_1 \ldots k_q}_{i_1 \ldots i_p} \left(A^{k_1 \ldots k_q}_{i_1 \ldots i_p} B^{k_1 \ldots k_q}_{i_1 \ldots i_p} \right)$$

in the basis $\boldsymbol{e}_i$.

For the definition of the operations of addition and subtraction of tensors to be correct, we must verify whether the coordinates of the sum (the difference) of the tensors are transformed by the law (8.19) of transformation of the coordinates of a tensor. To do that, we assume that alongside formulas (8.19) for the transformation of the coordinates of the tensor A we have analogous formulas for the transformation of the coordinates of the tensor B. Then by adding (subtracting) two formulas of this kind for the transformation of the coordinates of the tensors A and B we ascertain that the coordinates of the sum (the difference) are transformed by the law of transformation of the coordinates of a tensor.

2°. **Multiplication of a tensor by a number.** Suppose A is a tensor of type (p, q) having the coordinates $A^{k_1 \ldots k_q}_{i_1 \ldots i_p}$ in the basis $\boldsymbol{e}_i$ and α is an arbitrary real number.

The **product** αA **of the tensor** A **by the number** α is a tensor having the coordinates $\alpha A^{k_1 \ldots k_q}_{i_1 \ldots i_p}$ in the basis $\boldsymbol{e}_i$.

The fact that the coordinates $\alpha A^{k_1 \ldots k_q}_{i_1 \ldots i_p}$ are transformed by the law of tensors, can be seen from formulas (8.19).

3°. **Multiplication of tensors.** The operation of multiplication of tensors is defined for tensors of an arbitrary type.

Suppose A is a tensor of type (p, q) having coordinates $A^{k_1 \ldots k_q}_{i_1 \ldots i_p}$ in the given basis $\boldsymbol{e}_i$, and B is a tensor of type (r, s) having the coordinates $B^{m_1 \ldots m_s}_{l_1 \ldots l_r}$ in the same basis.

To define the **product** $D=AB$ **of the tensors** A and B, we form various products of the coordinates of the tensor A by the coordinates of the tensor B. In every such product, the indices $l_1, \ldots, l_r$ and $m_1, \ldots, m_s$ of the coordinates of the tensor B are replaced by new indices. Namely, we set $l_1 = i_{p+1}, \ldots, l_r = i_{p+r}$ and $m_1 = k_{q+1}, \ldots, m_s = k_{q+s}$.

The product $D=AB$ of the tensors A and B is a tensor of type $(p+r,\ q+s)$, having the coordinates

$$D^{k_1\ldots k_q\ k_{q+1}\ldots k_{q+s}}_{i_1\ldots i_p\ i_{p+1}\ldots i_{p+r}}=A^{k_1\ldots k_q}_{i_1\ldots i_p}B^{k_{q+1}\ldots k_{q+s}}_{i_{p+1}\ldots i_{p+r}} \quad (8.30)$$

in the basis $\boldsymbol{e}_i$.

To verify that the coordinates $D^{k_1\ldots k_q k_{q+1}\ldots k_{q+s}}_{i_1\ldots i_p i_{p+1}\ldots i_{p+r}}$ defined by relation (8.30) are transformed by the law of transformation of the coordinates of a tensor when we pass to a new basis, i.e. that they are indeed the coordinates of a tensor, it is sufficient to write the transformation formulas (8.19) for the coordinates $A^{k_1\ldots k_q}_{i_1\ldots i_p}$ and $B^{k_{q+1}\ldots k_{q+s}}_{i_{p+1}\ldots i_{p+r}}$ of the tensors A and B and multiply the right-hand and left-hand sides of the formulas. As a result, using relations (8.30), we can easily obtain the necessary formulas for the transformation of the coordinates $D^{k_1\ldots k_q k_{q+1}\ldots k_{q+s}}_{i_1\ldots i_p i_{p+1}\ldots i_{p+r}}$.

Remark. The operation of multiplication of tensors does not possess the commutative property: in general $AB\neq BA$. This can be explained by the fact that the sequence of the indices in the coordinates of a tensor defines the "number" of that coordinate.

Thus, although the numerical value of the expressions

$$A^{k_1\ldots k_q}_{i_1\ldots i_p}\ B^{k_{q+1}\ldots k_{q+s}}_{i_{p+1}\ldots i_{p+r}} \quad \text{and} \quad B^{k_{q+1}\ldots k_{q+s}}_{i_{p+1}\ldots i_{p+r}}\ A^{k_1\ldots k_q}_{i_1\ldots i_p}$$

is the same, the sequence of indices in these expressions is different and, therefore, they correspond to the coordinates with different "numbers". And this means that $AB\neq BA$.

4°. **Contraction of a tensor.** The operation of **contraction** is applied to a tensor of type $(p,\ q)$ in which $p\neq 0$ and $q\neq 0$ (i.e. to a tensor which has at least one superindex and one subscript).

Let A be a tensor of the indicated type. We shall describe the operation of contraction.

The contraction of the tensor A is performed by some marked superscript and subscript. As a result of contraction, a tensor of type $(p-1,\ q-1)$ is obtained.

Assume, for instance, that a superscript with the number m and a subscript with the number n are marked in each of the coordinates of the tensor A:

$$A^{k_1\ldots k_m\ldots k_q}_{i_1\ldots i_n\ldots i_p}.$$

Let us sum (contract) the coordinates of the tensor with similar marked indices. This operation yields the following coordinates of the tensor obtained by a contraction with respect to the superscript

and the subscript with the numbers m and n respectively:

$$A^{k_1 \ldots \alpha \ldots k_q}_{i_1 \ldots \alpha \ldots i_p} \tag{8.31}$$

(in expression (8.31) we have used the agreement on summation).

Let us verify that quantities (8.31) indeed form the coordinates of a tensor of type $(p-1,\ q-1)$. For that purpose we direct our attention to formulas (8.19). We rewrite these formulas in the following form, having marked the indices of interest to us (we underline these indices by brackets):

$$A^{k'_1 \ldots \underbrace{k'_m} \ldots k'_q}_{i'_1 \ldots \underbrace{i'_n} \ldots i'_p} = b^{k_1}_{k'_1} \ldots b^{k_m}_{k'_m} \ldots b^{k_q}_{k'_q} b^{i_1}_{i'_1} \ldots b^{i_n}_{\underbrace{i'_n}} \ldots b^{i_p}_{i'_p} A^{k_1 \ldots k_m \ldots k_q}_{i_1 \ldots i_n \ldots i_p}. \tag{8.32}$$

Let us perform a summation in the right-hand and left-hand sides of (8.32) with respect to the emphasized indices k'_m and i'_n. To do that, it is sufficient to set these indices equal to α' and make use of the agreement on summation. As a result, we get a certain equality. We find the expressions on the right-hand and left-hand sides of this equality. On the left-hand side we evidently get an expression

$$A^{k'_1 \ldots \alpha' \ldots k'_q}_{i'_1 \ldots \alpha' \ldots i'_p} . \tag{8.33}$$

On the right-hand side the product of $b^{\alpha'}_{k_m}$ by $b^{i_n}_{\alpha'}$ is equal to $\delta^{i_n}_{k_m}$, i.e. equal to unity for $k_m = i_n = \alpha$ and to zero for $k_m \neq i_n$. Thus, on the right-hand side we get an expression

$$b^{k'_1}_{k_1} \ldots b^{k'_q}_{k_q} b^{i_1}_{i'_1} \ldots b^{i_p}_{i'_p} A^{k_1 \ldots \alpha \ldots k_q}_{i_1 \ldots \alpha \ldots i_p} . \tag{8.34}$$

Comparing expressions (8.33) and (8.34), we make sure that the quantities $A^{k_1 \ldots \alpha \ldots k_q}_{i_1 \ldots \alpha \ldots i_p}$ are transformed in passing to a new basis by the law of transformation of the coordinates of a tensor. This tensor is evidently of the type $(p-1,\ q-1)$.

Remark. The term "contraction of tensors" can also be used in the following sense.

Let us consider two tensors A and B. The coordinates of the first tensor have at least one superscript k and the coordinates of the second have at least one subscript i.

We form a product AB of these tensors and then contract the tensor AB with respect to the superscript k and the subscript i. The terminology usually used for this operation is the "*contraction of tensors A and B with respect to the indices k and i*".

5°. **Commutation of indices.** This operation consists in subjecting the indices of each coordinate of a tensor to the same commutation in

any basis. This means that we "enumerate" the coordinates of a given tensor in a different way. The reader is invited to verify that a tensor (in general different from the given tensor) results.

6°. **Symmetrization and alternation.** We must first introduce the concepts of a **symmetric** and a **skew-symmetric** tensor.

The tensor A with the coordinates

$$A^{k_1 \ldots k_q}_{i_1 \ldots i_m \ldots i_n \ldots i_p} \tag{8.35}$$

is said to be **symmetric** with respect to the subscripts i_m and i_n if as a result of the commutation of these indices* the coordinates of the tensor A do not change their meaning, i.e.

$$A^{k_1 \ldots k_q}_{i_1 \ldots i_m \ldots i_n \ldots i_p} = A^{k_1 \ldots k_q}_{i_1 \ldots i_n \ldots i_m \ldots i_p}. \tag{8.36}$$

Relation (8.36) is called the **condition of symmetry** of the tensor A with respect to subscripts labelled m and n.

The tensor A is said to be **skew-symmetric** with respect to the subscripts i_m and i_n if upon an interchange of these indices the following relation holds true:

$$A^{k_1 \ldots k_q}_{i_1 \ldots i_m \ldots i_n \ldots i_p} = -A^{k_1 \ldots k_q}_{i_1 \ldots i_n \ldots i_m \ldots i_p}. \tag{8.37}$$

Relation (8.37) is called the **condition of skew-symmetry** of the tensor A with respect to the subscripts labelled m and n.

The concepts of symmetry and skew-symmetry of a tensor with respect to two superscripts can be introduced by analogy.

Remark. If the condition of symmetry (skew-symmetry) with respect to two subscripts i_m and i_n is fulfilled for the tensor A in a given system of coordinates, then it is also fulfilled in any other system of coordinates.

We shall now describe the **operation of symmetrization.**

Let A be a tensor of type (p, q) with coordinates (8.35). We interchange the subscripts with the numbers m and n in each coordinate and construct a tensor $A_{(m,\, n)}$ with coordinates

$$\frac{1}{2}\left(A^{k_1 \ldots k_q}_{i_1 \ldots i_m \ldots i_n \ldots i_p} + A^{k_1 \ldots k_q}_{i_1 \ldots i_n \ldots i_m \ldots i_p}\right). \tag{8.38}$$

The operation of constructing a tensor $A_{(m,\, n)}$ is called the **operation of symmetrization** of the tensor A with respect to the subscripts labelled m and n.

Note that coordinates (8.38) of the tensor $A_{(m,\, n)}$ are usually sym-

* Recall that when we commute the indices of the coordinates of a tensor, we obtain in general coordinates of another tensor.

bolized as

$$A^{k_1 \ldots k_q}_{i_1 \ldots (i_m \ldots i_n) \ldots i_p}. \tag{8.39}$$

The condition of symmetry (8.36) with respect to the subscripts labelled m and n is evidently fulfilled for the tensor $A_{(m,\ n)}$.

The operation of symmetrization of a tensor **with respect to the superscripts** labelled m and n can be defined by analogy. The tensor that is constructed is symbolized as $A^{(m,\ n)}$. The designation used for the coordinates of the tensor $A^{(m,\ n)}$ is similar to (8.39) for the coordinates of the tensor $A_{(m,\ n)}$.

The **operation of alternating** the tensor A with respect to the subscripts labelled m and n is carried out as follows.

The subscripts labelled m and n of each coordinate of the tensor A are interchanged and then a tensor $A_{(m,\ n)}$ with coordinates

$$\frac{1}{2}\left(A^{k_1 \ldots k_q}_{i_1 \ldots i_m \ldots i_n \ldots i_p} - A^{k_1 \ldots k_q}_{i_1 \ldots i_n \ldots i_m \ldots i_p}\right). \tag{8.40}$$

is constructed.

The operation of constructing a tensor $A_{(m,\ n)}$ is called the **operation of alternation** of the tensor A with respect to the subscripts labelled m and n. Coordinates (8.40) of the tensor $A_{[m,\ n]}$ are usually symbolized as

$$A^{k_1 \ldots k_q}_{i_1 \ldots [i_m \ldots i_n] \ldots i_p}. \tag{8.41}$$

Condition (8.37) of skew-symmetry with respect to the subscripts i_m and i_n is evidently fulfilled for the tensor $A_{[m,\ n]}$.

The operation of alternating a tensor **with respect to the superscripts** i_m and i_n can be carried out by analogy. The constructed tensor is symbolized as $A^{[m,\ n]}$. The designation similar to (8.41) for the coordinates of the tensor $A_{[m,\ n]}$ is used for the coordinates of the tensor $A^{[m,\ n]}$.

We must point out in conclusion the obvious equality

$$A = A_{(m,\ n)} + A_{[m,\ n]}.$$

8.3. A Metric Tensor. Principal Operations of Vector Algebra in Tensor Designations

8.3.1. The concept of a metric tensor in a Euclidean space. We mentioned in 7.2 that in a finite-dimensional linear space a scalar product could be defined by means of a bilinear form which is a polar of some positive-definite quadratic form. We assume in this section that

in the finite-dimensional Euclidean space E^n the scalar product is defined by this type of a bilinear form $A(\boldsymbol{x}, \boldsymbol{y})$.

We have ascertained in 8.2.2 (Example 3) that the coefficients of the matrix of a bilinear form can be regarded as the coordinates of a tensor. For the bilinear form $A(\boldsymbol{x}, \boldsymbol{y})$ which defines the scalar product in E^n, we shall designate these coefficients as g_{ij}. Thus g_{ij} are the coordinates of the tensor G in the basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$. This tensor of type (2, 0) is called a **metric tensor** of the space E^n. Recall that the coordinates g_{ij} of the tensor G are defined by the relations

$$g_{ij} = A(\boldsymbol{e}_i, \boldsymbol{e}_j) \tag{8.42}$$

(see formula (8.23)).

It should also be pointed out that since the form $A(\boldsymbol{x}, \boldsymbol{y})$ is symmetric ($A(\boldsymbol{x}, \boldsymbol{y}) = A(\boldsymbol{y}, \boldsymbol{x})$), it follows, according to (8.42), that $g_{ij} = g_{ji}$, i.e. the *metric tensor G is symmetric with respect to the subscripts i and j.*

Suppose that $\boldsymbol{x}$ and $\boldsymbol{y}$ are arbitrary vectors in E^n, and x^i and y^j are the coordinates of these vectors in the basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$.

The scalar product $(\boldsymbol{x}, \boldsymbol{y})$ of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is equal to $A(\boldsymbol{x}, \boldsymbol{y})$. Directing our attention to expression (8.24) for a bilinear form in the given basis and using the equation $(\boldsymbol{x}, \boldsymbol{y}) = A(\boldsymbol{x}, \boldsymbol{y})$, we get the following formula for the scalar product $(\boldsymbol{x}, \boldsymbol{y})$ of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$

$$(\boldsymbol{x}, \boldsymbol{y}) = g_{ij} x^i y^j. \tag{8.43}$$

In particular, the scalar products $(\boldsymbol{e}_i, \boldsymbol{e}_j)$ of the base vectors $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ are equal to g_{ij}:

$$(\boldsymbol{e}_i, \boldsymbol{e}_j) = g_{ij} \tag{8.44}$$

(this follows from the equality $(\boldsymbol{e}_i, \boldsymbol{e}_j) = A(\boldsymbol{e}_i, \boldsymbol{e}_j)$ and from formula (8.42); by the way, formula (8.44) can be easily obtained directly from (8.43)).

Let us now consider a reciprocal basis $\boldsymbol{e}^1, \boldsymbol{e}^2, \ldots, \boldsymbol{e}^n$ alongside the basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$. Let $\boldsymbol{x} = x_j \boldsymbol{e}^i$ and $\boldsymbol{y} = y_j \boldsymbol{e}^j$ be representations of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in terms of vectors of the reciprocal basis. Then we get the following formula for the scalar product $(\boldsymbol{x}, \boldsymbol{y})$:

$$(\boldsymbol{x}, \boldsymbol{y}) = A(\boldsymbol{x}, \boldsymbol{y}) = A(x_i \boldsymbol{e}^i, y_j \boldsymbol{e}^j) = A(\boldsymbol{e}^i, \boldsymbol{e}^j) x_i y_j.$$

We introduce the designation

$$g^{ij} = A(\boldsymbol{e}^i, \boldsymbol{e}^j) \tag{8.45}$$

and get the following expression for $(\boldsymbol{x}, \boldsymbol{y})$:

$$(\boldsymbol{x}, \boldsymbol{y}) = g^{ij} x_i y_j. \tag{8.46}$$

As in 8.2.2 (see Example 3), it is easy to verify that g^{ij} are the coordinates of a tensor of type (0, 2) which is symmetric with respec

to the indices i and j. This tensor of type (0, 2) is also called a **metric tensor** of the space E^n.

We denote it by the same symbol G as the metric tensor of type (2, 0) introduced above: in 3.2 we shall find that the coordinates g_{ij} and g^{ij} can be regarded as the covariant and contravariant coordinates of the same vector. In what follows, we shall call the coordinates g_{ij} and g^{ij} the covariant and contravariant coordinates of the tensor G.

At the end of 8.1.2, while investigating the problem of constructing reciprocal bases, we introduced the quantities g_{ij} and g^{ij} by formulas (8.10). Comparing these formulas with (8.42) and (8.45), we infer that these quantities are the covariant and contravariant coordinates of the metric tensor G.

We have also proved in 8.1.2 that the matrices whose elements are the coordinates g_{ij} and g^{ij}, are mutually inverse. This means that there holds a relation

$$g^{ij}g_{jk}=\delta^i_k. \tag{8.47}$$

Thus the coordinates g^{ij} of the tensor G can be constructed from the coordinates g_{ij} and vice versa (for that purpose, we must use the familiar method of constructing the elements of an inverse matrix).

8.3.2. The operations of lifting and lowering indices with the aid of a metric tensor. The metric tensor G is used for the **lifting** and **lowering of indices** of the coordinates of a given tensor A.

The operation consists in the following.

Let A be a tensor of type (p, q) with coordinates $A^{k_1k_2\ldots k_q}_{i_1i_2\ldots i_p}$. We shall describe, for example, an operation of lifting the index i_1. We contract the tensors G and A with respect to the superscript j in the first tensor and with respect to the subscript i_1 of the second tensor, i.e. construct a tensor with coordinates $g^{i\alpha}\,A^{k_1k_2\ldots k_q}_{\alpha i_2\ldots i_p}$ and designate the index i in the coordinates of the tensor obtained as i_1. Then we use the symbol $A^{i_1k_1k_2\ldots k_q}_{i_2\ldots i_p}$ for these coordinates. Thus we have

$$A^{i_1k_1k_2\ldots k_q}_{i_2\ldots i_p}=g^{i_1\alpha}\,A^{k_1k_2\ldots k_q}_{\alpha i_2\ldots i_p} \tag{8.48}$$

Remark 1. Since the sequence of indices in the coordinates of a tensor determines the "*numeration*" of its coordinates, it follows, in a general case, that when we lift an index, we must mark the place, among the superscripts, to which the given subscript will be lifted. It is sometimes necessary to mark the place of the index being lifted among the subscripts. This is done with the aid of a dot which is put into the place of the lifted index. Therefore, the coordinates of the tensor on the left-hand side of (8.48) should have been written as $A^{i_1k_1k_2\ldots k_q}_{i_2\ldots i_p}$.

For example, if we lift the *first* subscript to the *second* place among the superscripts, we get a tensor with the coordinates $A^{k_1 i_i k_2 \ldots k_q}_{i_2 \ldots i_q}$.

Remark 2. The operation of lowering an index with the aid of a fundamental tensor G can be defined by analogy. For instance, the coordinates of a tensor obtained by lowering the index k_q of the tensor A to the last place among the subscripts have the following form:

$$A^{k_1 \ldots k_{q-1}}_{i_1 \ldots i_p k_q} = g_{k_q \alpha} \; A^{k_1 \ldots k_{q-1} \alpha}_{i_1 \ldots i_p} .$$

Remark 3. The operation of lifting or lowering an index can be performed several times, each time for a different index of a given tensor.

Let us consider some examples of lifting and lowering indices of tensors.

Assume that $\boldsymbol{x}$ is a vector, and x_i and x^i are, respectively, its covariant and contravariant coordinates (recall that a vector is a tensor of rank 1).

Let us lift the index i in the coordinates x_i by means of the fundamental tensor G. As a result we get a tensor with the coordinates $g^{i\alpha} x_\alpha$. Since $x_\alpha = (\boldsymbol{x}, \boldsymbol{e}_\alpha)$ we have $g^{i\alpha} x_\alpha = g^{i\alpha} (\boldsymbol{x}, \boldsymbol{e}_\alpha) = (x, g^{i\alpha} \boldsymbol{e}_\alpha)$.

According to (8.11), $g^{i\alpha} \boldsymbol{e}_\alpha = \boldsymbol{e}^i$ and $(\boldsymbol{x}, \boldsymbol{e}^i) = x^i$. Therefore, $g^{i\alpha} x_\alpha = x^i$.

Thus the contravariant coordinates x^i of the vector $\boldsymbol{x}$ can be obtained as a result of an operation of lifting an index of the covariant coordinates x_i of that vector.

The covariant coordinates x_i can be obtained as a result of an operation of lowering an index of the contravariant coordinates x^i.

Let us find the result of an operation of lifting an index applied two times to the covariant coordinates g_{ij} of metric tensor G with the aid of the contravariant coordinates g^{ij} of that tensor. In other words, we shall find what the tensor with the coordinates

$$g^{i\alpha} g^{j\beta} g_{\alpha\beta} \tag{8.49}$$

is like.

Using the symmetry of the tensor G with respect to the subscripts and relation (8.47), we find that $g^{j\beta} g_{\alpha\beta} = g^{j\beta} g_{\beta\alpha} = \delta^i_\alpha$. Substituting the expression found for $g^{j\beta} g_{\alpha\beta}$ into (8.49) and using the properties of the Kronecker delta δ^j_α, we get

$$g^{i\alpha} g^{j\beta} g_{\alpha\beta} = g^{ij}.$$

By a complete analogy we can verify the validity of the equality

$$g_{i\alpha} g_{j\beta} g^{\alpha\beta} = g_{ij}.$$

The last two formulas emphasize once again that it is natural to

regard g_{ij} and g^{ij} as covariant and contravariant coordinates of the metric tensor G.

8.3.3. Orthonormal bases in E^n. We have elucidated that the scalar product $(\boldsymbol{x}, \boldsymbol{y})$ in E^n can be defined by the metric tensor G whose coordinates g_{ij} are the elements of the symmetric positive definite matrix (g_{ij}). Namely, according to (8.43),

$$(\boldsymbol{x}, \boldsymbol{y}) = g_{ij}x^i y^j.$$

It is known that by means of a transformation of the basis, the matrix of the bilinear form $g_{ij}x^i y^j$ can be reduced to a diagonal form. In this case, by virtue of the positive definiteness of the matrix, after the reduction of the matrix (g_{ij}) to a diagonal form, the coordinates of the metric tensor are equal to zero for $i \neq j$ and to unity for $i = j$. We symbolize these coordinates as g_{ij} and obtain

$$g_{ij} = \begin{cases} 0 & \text{for} \quad i \neq j, \\ 1 & \text{for} \quad i = j. \end{cases} \tag{8.50}$$

The basis $\boldsymbol{e}_i$ in which the coordinates g_{ij} of the metric tensor satisfy condition (8.50) is orthonormal. Indeed, since $(\boldsymbol{e}_i, \boldsymbol{e}_j) = g_{ij}$ (see (8.44)), we get, according to (8.50),

$$(\boldsymbol{e}_i, \boldsymbol{e}_j) = \begin{cases} 0 \text{ for } i \neq j, \\ 1 \text{ for } i = j, \end{cases}$$

and this means that $\boldsymbol{e}_i$ is an orthonormal basis.

We found in Chapter 4 that the scalar product $(\boldsymbol{x}, \boldsymbol{y})$ of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ with the coordinates x^i and y^i could be calculated in an orthonormal basis by the formula

$$(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} x^i y^i, \tag{8.51}$$

and the square of the length $(\boldsymbol{x}, \boldsymbol{x})$ of the vector $\boldsymbol{x}$ by the formula

$$(\boldsymbol{x}, \boldsymbol{x}) = \sum_{i=1}^{n} (x^i)^2. \tag{8.52}$$

Let us direct our attention to the so-called ***orthogonal*** *linear transformations,* i.e. linear transformations under which an orthonormal basis passes into an orthonormal one. In other words, if L is an orthogonal transformation and $\boldsymbol{e}_i$ is an orthonormal basis, the $L\boldsymbol{e}_i$ also forms an orthonormal basis.

Let us see how the transformation L acts on an arbitrary vector $\boldsymbol{x} = x^i \boldsymbol{e}_i$. We denote by $\boldsymbol{X}$ the result of the action of L on $\boldsymbol{x}$:

$$\boldsymbol{X} = L\boldsymbol{x}.$$

Using the property of linearity of L, we find that

$$X = Lx^i e_i = x^i L e_i.$$

Since Le_i is a basis, it follows from the last relation that the vector X has the same coordinates in the basis Le_i as the vector x in the basis e_i, i.e. the coordinates of the vector retain their values under an orthogonal transformation.

Since Le_i is an orthonormal basis, the scalar product (X, Y) of the vectors $X = Lx$ and $Y = Ly$ can be found by formula (8.51), and the square of the length (X, X) of the vector $X = Lx$ by formula (8.52). We have found that the coordinates of a vector retain their value under an orthogonal transformation. From this and from relations (8.51) and (8.52) we obtain

$$(X, Y) = (x, y), \ (X, X) = (x, x).$$

Thus *the lengths of vectors and their scalar products do not change under an orthogonal transformation.*

As is known, the orthogonal transformations L can be defined by means of an orthogonal matrix. The determinant $\det L$ of such a matrix satisfies the condition $\det L = \pm 1$.

Let us choose one of the orthonormal bases and agree to call it a *right* basis. In that case we say that the Euclidean space E^n is *oriented*. All the bases in E^n which are obtained from the given basis by means of orthogonal transformations with a determinant equal to $+1$ are called *right* bases and all bases which result from the given basis by means of orthogonal transformations with the determinant equal to -1 are called *left* bases. It is easy to verify when a right basis is transformed into a right one the determinant of the transformation is equal to $+1$ and when a left basis is transformed into a left one the determinant is equal to -1.

We denote by $O(n)$ the set of all orthogonal transformations in E^n and by $O_+(n)$ the set of all orthogonal transformations of right bases.

We shall consider these sets in the next chapter.

Remark. In what follows, we call the arbitrary basis $e_1, e_2, \ldots, e_n$ the *right basis* (the *left* basis) if the determinant of the transition matrix from the chosen orthonormal basis to the basis $e_1, e_2, \ldots, e_n$ is positive (negative).

8.3.4. A discriminatory tensor. Let us consider the so-called **well-skew-symmetric tensor** $\varepsilon_{i_1 i_2 \ldots i_p}$ of type $(p, 0)$, i.e. a tensor which is skew-symmetric with respect to any two subscripts.

For this tensor to be nonzero, it is necessary that the number p should not exceed n, i.e. satisfy the condition $p \leqslant n$ since in the case of $p > n$ any coordinate has at least two identical indices upon an interchange of which this coordinate must simultaneously change sign and remain unchanged. And this is only possible when this

coordinate is equal to zero. Consequently, when $p > n$, any coordinate of a tensor is equal to zero, i.e. the tensor is zero.

Of special interest is a well-skew-symmetric tensor whose rank p is equal to the dimension n of the space.

Any coordinate $\varepsilon_{i_1 i_2 \ldots i_n}$ of such a tensor can be found by the formula

$$\varepsilon_{i_1 i_2 \ldots i_n} = \begin{cases} 0, \text{ if at least two of the indices } i_1, i_2, \ldots i_n, \text{coincide,} \\ (-1)^{\operatorname{sign} \sigma} \varepsilon_{12 \ldots n} \text{ if all the indices are different.} \end{cases} \tag{8.53}$$

In formula (8.53) sign σ is equal to 0 or $+1$ according as the interchange $\sigma = (i_1, i_2, \ldots, i_n)$ is even or odd (sign σ is also called the *sign* of that interchange).

Let us consider some right orthonormal system of coordinates and set

$$\varepsilon_{12 \ldots n} = 1 \tag{8.54}$$

in it. In this coordinate system relation (8.53) defines all the coordinates $\varepsilon_{i_1 i_2 \ldots i_n}$ of a well-skew-symmetric tensor and, consequently, the tensor itself. In what follows we shall call this tensor a **discriminatory tensor.** We denote the coordinates of this tensor in the arbitrary basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ by $c_{i_1 i_2 \ldots i_n}$.

We shall use the symbol B for the transition matrix from the chosen right orthonormal basis to some basis $\boldsymbol{e}_{1'}, \boldsymbol{e}_{2'}, \ldots, \boldsymbol{e}_{n'}$ and the symbol $b^i_{i'}$ for the elements of that matrix.

In accordance with (8.53), to calculate the coordinates $c_{i'_1 i'_2 \ldots i'_n}$ of the discriminatory tensor in the basis $\boldsymbol{e}_{1'}, \boldsymbol{e}_{2'}, \ldots, \boldsymbol{e}_{n'}$ it is sufficient to know the value of the coordinate $c_{1'2' \ldots n'}$.

Using formula (8.19) for the transformation of the coordinates of a tensor and relation (8.54), and passing from the chosen orthonormal basis to the basis $\boldsymbol{e}_{1'}, \boldsymbol{e}_{2'}, \ldots, \boldsymbol{e}_{n'}$ we get

$$c_{1'2' \ldots n'} = b^{i_1}_{1'} b^{i_2}_{2'} \ldots b^{i_n}_{n'} \varepsilon_{i_1 i_2 \ldots i_n}$$

$$= \varepsilon_{12 \ldots n} \sum_{\sigma = (i_1, i_2, \ldots, i_n)} (-1)^{\operatorname{sign} \sigma} b^{i_1}_{1'} b^{i_2}_{2'} \ldots b^{i_n}_{n'} = \det (b^i_{i'}) = \det B. \tag{8.55}$$

Assume that $g_{i'j'}$ are the coordinates of the fundamental tensor in basis $\boldsymbol{e}_{1'}, \boldsymbol{e}_{2'}, \ldots, \boldsymbol{e}_{n'}$. Since $\tilde{G} = (g_{i'j'})$ is the matrix of the bilinear form $g_{i'j'} x^{i'} x^{j'}$, which is the scalar product of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ with the coordinates $x^{i'}$ and $y^{i'}$, the formula $\tilde{G} = B'IB$ is valid when we pass from the given orthonormal basis (in which the matrix I of the bilinear form being considered is an identity matrix) to the basis

$e_{1'}, e_{2'}, \ldots, e_{n'}$. Hence it follows that $\det \tilde{G} = \det B' \det I \det B = (\det B)^2$.

Designating $\det \tilde{G}$ as g, we get $\det B = \pm\sqrt{g}$ from the last relation. Using relations (8.55), we find that $c_{1'2'\ldots n'} = \pm\sqrt{g}$. Thus, in the arbitrary basis $e_1, e_2, \ldots, e_n$ the expression for the coordinate $c_{12\ldots n}$ of the discriminatory tensor has the form

$$c_{12\ldots n} = \pm\sqrt{g}, \tag{8.56}$$

where g is the determinant of the matrix (g_{ij}) of the metric tensor in the basis $e_1, e_2, \ldots, e_n$.

Note that the plus sign in formula (8.56) corresponds to the right basis and the minus sign to the left basis.

8.3.5. An oriented volume. Let us introduce, in the oriented Euclidean space E^n, the so-called **affine system of coordinates** defining it as the set of a fixed point O with the coordinates $(0, 0, \ldots, 0)$ and a basis $e_1, e_2, \ldots, e_n$. The coordinates of any point M in E^n are defined in this case as the coordinates in the basis $e_1, e_2, \ldots, e_n$ of the vector $\overrightarrow{OM}$.

Let us consider in E^n a numbered system of n vectors

$$\overset{1}{x}, \overset{2}{x}, \ldots, \overset{n}{x} \tag{8.57}$$

and various vectors $\overrightarrow{OM}$ defined by the relations

$$\overrightarrow{OM} = \alpha_1 \overset{1}{x} + + \alpha_2 \overset{2}{x} + \ldots \alpha_n \overset{n}{x}, \tag{8.58}$$

for various α_i satisfying the inequalities $0 \leqslant \alpha_i \leqslant 1$, $i = 1, 2, \ldots, n$.

The set of all points M of the space E^n defined by relations (8.58) forms the so-called ***n*-dimensional parallelepiped** in E^n which is spanned by vectors (8.57).

The **oriented volume** $V(\overset{1}{x}, \overset{2}{x}, \ldots, \overset{n}{x})$ of that parallelepiped is the number

$$V(\overset{1}{x}, \overset{2}{x}, \ldots, \overset{n}{x}) = c_{i_1 i_2 \ldots i_n} \overset{1}{x}{}^{i_1} \overset{2}{x}{}^{i_2} \ldots, \overset{n}{x}{}^{i_n}. \tag{8.59}$$

In this case, $c_{i_1 i_2 \ldots i_n}$ are the coordinates of the discriminatory tensor in the basis $e_1, e_2, \ldots, e_n$, and $\overset{1}{x}{}^{i_1}, \overset{2}{x}{}^{i_2}, \ldots, \overset{n}{x}{}^{i_n}$ are the contravariant coordinates of the vectors $\overset{1}{x}, \overset{2}{x}, \ldots, \overset{n}{x}$ in that basis.

We use the term "oriented volume" since in the case when vectors (8.57) form a right basis the oriented volume is positive ($V > 0$) and in the case of a left basis it is negative ($V < 0$).

Note that for $n = 3$ the oriented volume calculated for $n = 3$ by formula (8.59) is an ordinary volume of a parallelepiped spanned by the vectors $\overset{1}{x}, \overset{2}{x}, \overset{3}{x}$. The volume is taken with the plus sign if the triad

$\overset{1}{\boldsymbol{x}}, \overset{2}{\boldsymbol{x}}, \overset{3}{\boldsymbol{x}}$ is right-handed and with the minus sign if the triad is left-handed.

8.3.6. A vector product. We can use a discriminatory tensor to write a vector product in the three-dimensional space E^3 in tensor form. Such a notation is widely used in various calculations in the so-called curvilinear coordinates.

Suppose c_{ijk} are the coordinates of a discriminatory tensor in the given basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$ of the space E^3. Let us lift the first index i of this tensor with the aid of the metric tensor g_{lm}, i.e. consider the tensor c^i_{jk}. Then the coordinates z^i of the vector $z=[\boldsymbol{xy}]$ (i.e. of the vector product of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$) in the basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$ have the form

$$z^i = c^i_{jk} x^j y^k. \tag{8.60}$$

Since $c^i_{jk} x^j y^k$ is a tensor of type (0, 1), z^i can be regarded as the contravariant coordinates of the vector. Therefore, to ascertain that z^i are indeed the coordinates of a vector product, it is sufficient to carry out the verification in some definite system of coordinates. This verification is elementary for an orthonormal basis and we invite the reader to carry it out.

Relation (8.60) can serve as the basis for the introduction of the vector product of $n-1$ vectors in E^n.

Suppose $\overset{1}{\boldsymbol{x}}, \overset{2}{\boldsymbol{x}} \ldots, \overset{n-1}{\boldsymbol{x}}$ are some $n-1$ vectors in E^n. Let us find the coordinates z^i of the vector product $\boldsymbol{z} = \left[\overset{1}{\boldsymbol{x}}\overset{2}{\boldsymbol{x}} \ldots \overset{n-1}{\boldsymbol{x}}\right]$ by means of the relations

$$z^i = c^i_{i_1 i_2 \ldots i_{n-1}} \overset{1}{\boldsymbol{x}}{}^{i_1} \overset{2}{\boldsymbol{x}}{}^{i_2} \ldots \overset{n-1}{\boldsymbol{x}}{}^{i_{n-1}}. \tag{8.61}$$

In relations (8.61) $c^i_{i_1 i_2 \ldots i_{n-1}}$ are the coordinates of a descriminatory tensor with the lifted first index, and $\overset{1}{x}{}^{i_1}, \overset{2}{x}{}^{i_2}, \ldots, \overset{n-1}{x}{}^{i_{n-1}}$ are the contravariant coordinates of the vectors $\overset{1}{\boldsymbol{x}}, \overset{2}{\boldsymbol{x}}, \ldots, \overset{n-1}{\boldsymbol{x}}$.

8.3.7. A triple vector product. The reader is sure to be familiar, from vector algebra, with the following formula for the triple vector product $[\boldsymbol{a}[\boldsymbol{bd}]]$ of the vectors $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{d}$:

$$[\boldsymbol{a}[\boldsymbol{bd}]] = \boldsymbol{b}(\boldsymbol{ad}) - \boldsymbol{d}(\boldsymbol{ab}). \tag{8.62}$$

Using relation (8.60) and formula (8.43) for the scalar product of vectors rewrite (8.62) as follows:

$$c^i_{kl} a^k c^l_{mn} b^m d^n = b^i g_{kl} a^k d^l - d^i g_{kl} a^k b^l. \tag{8.63}$$

With the aid of (8.63), we get a formula relating the tensors c^i_{kl} and g_{ik}, which we shall use for writing the coordinates of the triple vector product.

Let us perform the following transformations in formula (8.63). In

the first term $b^i g_{kl}\, a^k d^l$ on the right-hand side of (8.63) we replace b^i by $b^m \delta^i_m$* and replace the sum index l by n. In the second term on the right-hand side of (8.63) we set $d^i = d^n \delta^i_n$ and replace the sum index l by m. After these transformations, formula (8.63) assumes the form

$$(c^i_{kl} c^l_{mn})\, a^k b^m d^n = (g_{kn} \delta^i_m - g_{km} \delta^i_n)\, a^k b^m d^n. \tag{8.64}$$

Since relation (8.64) is valid for any vectors $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{d}$, it is an identity with respect to the coordinates a^k, b^m and d^n of these vectors and, therefore, the equality

$$c^i_{kl} c^l_{mn} = g_{kn} \delta^i_m - g_{km} \delta^i_n \tag{8.65}$$

holds for any indices i, k, m and n. We designate the coordinates of the triple vector product $[\boldsymbol{a}[\boldsymbol{bd}]]$ as z^i. Then, according to (8.63), $z^i = c^i_{kl} c^l_{mn} a^k b^m d^n$. From this and from (8.65) we get the following expression for the coordinates z^i of the triple vector product $[\boldsymbol{a}[\boldsymbol{bd}]]$:

$$z^i = (g_{kn} \delta^i_m - g_{km} \delta^i_n)\, a^k b^m d^n. \tag{8.66}$$

Formula (8.66) is convenient for various applications.

8.4. A Metric Tensor of a Pseudo-Euclidean Space

8.4.1. The concept of a pseudo-Euclidean space and a metric tensor of a pseudo-Euclidean space. Let us consider an n-dimensional linear space L in which a nonsingular symmetric bilinear form $A(\boldsymbol{x}, \boldsymbol{y})$ is defined which is the polar of an alternating quadratic form.

The **scalar product** $(\boldsymbol{x}, \boldsymbol{y})$ of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is the value $A(\boldsymbol{x}, \boldsymbol{y})$ of the bilinear form. The name "scalar product" is conventional since in this case the fourth axiom of the scalar product does not hold true. Namely, in the case when the bilinear form $A(\boldsymbol{x}, \boldsymbol{y})$ is the polar of the alternating quadratic form, the expression $A(\boldsymbol{x}, \boldsymbol{x})$ can have either a positive or a negative value depending on the choice of $\boldsymbol{x}$ and vanishes for nonzero values of $\boldsymbol{x}$. All the same we shall use the term "scalar product" since it is customary.

Here is the definition of a pseudo-Euclidean space.

Definition. ***A pseudo-Euclidean space*** *is an n-dimensional linear space L in which a scalar product is defined by means of a nonsingular symmetric bilinear form $A(\boldsymbol{x}, \boldsymbol{y})$ which is the polar of an alternating quadratic form.*

The number n is the **dimension** of the pseudo-Euclidean space.

Let us isolate a basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n$ in the linear space L and denote by (g_{ij}) the matrix of the bilinear form $A(\boldsymbol{x}, \boldsymbol{y})$ in that basis (recall that $g_{ij} = A(\boldsymbol{e}_i, \boldsymbol{e}_j)$). If x^i and y^j are the contravariant coordi-

* The relation $b' = b^m \delta^i_m$ follows from the properties of the Kronecker delta.

nates of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, then

$$A(\boldsymbol{x}, \boldsymbol{y}) = g_{ij}x^i y^j. \tag{8.67}$$

By a complete analogy with the arguments presented in 8.2.2 we can prove that g_{ij} are the coordinates of the tensor G of type (2, 0). We shall call this tensor a **metric tensor of a pseudo-Euclidean space.**

Since the scalar product $(\boldsymbol{x}, \boldsymbol{y})$ is equal to $A(\boldsymbol{x}, \boldsymbol{y})$, we have, in accordance with (8.67), $(\boldsymbol{x}, \boldsymbol{y}) = g_{ij}x^i y^j$.

It is known that the matrix (g_{ij}) of the bilinear form $A(\boldsymbol{x}, \boldsymbol{y})$ can be reduced to a diagonal form. In that case, since the form $A(\boldsymbol{x}, \boldsymbol{y})$ is nonsingular, the coordinates g_{ij} of the metric tensor are equal to zero for $i \neq j$ after the reduction to a diagonal form, and to unity or minus unity for $i = j$. The number p of positive and the number q of negative diagonal elements does not depend on the way of reducing the matrix to a diagonal form, and $p + q = n$ since the form $A(\boldsymbol{x}, \boldsymbol{y})$ is nonsingular.

These arguments explain the designation $E^n_{(p, q)}$ for an n-dimensional pseudo-Euclidean space.

It is natural to pose a question concerning the measurement of the lengths of vectors in a pseudo-Euclidean space.

In a Euclidean space with a metric tensor g_{ij} the square of the length of the vector $\boldsymbol{x}$ with the coordinates x^i is assumed to be equal to $g_{ij}x^i x^j$. If we define the square of the length $s^2(\boldsymbol{x})$ of the vector $\boldsymbol{x}$ by means of the relation

$$s^2(\boldsymbol{x}) = g_{ij}x^i x^j, \tag{8.68}$$

then, evidently (since the form $A(\boldsymbol{x}, \boldsymbol{x})$ is alternating), there are nonzero vectors with a positive square of length, with a negative square of length and with a zero square of length. Therefore, to obtain only real numbers as a measure of the length of vectors, the relation

$$\sigma(\boldsymbol{x}) = (\operatorname{sgn} s^2(\boldsymbol{x}))\sqrt{|s^2(\boldsymbol{x})|} \tag{8.69}$$

is usually taken as the length of a vector.

In what follows we shall use the following terminology borrowed from the special theory of relativity: we shall call the nonzero vector $\boldsymbol{x}$ **time-like** if $\sigma(\boldsymbol{x}) > 0$ for that vector, **space-like** if $\sigma(\boldsymbol{x}) < 0$, and **isotropic** if $\sigma(\boldsymbol{x}) = 0$.

The following statement holds true.

The set of termini of all time-like (space-like, isotropic) vectors whose origins coincide with the arbitrary fixed point M of the pseudo-Euclidean space, forms a cone.

For definiteness, we shall prove the statement for the case of time-like vectors. It is evidently sufficient to prove that if $\boldsymbol{x}$ is a time-like vector, then the vector $\lambda\boldsymbol{x}$ is also time-like for any real $\lambda \neq 0$.

Since the coordinates of the vector $\lambda \boldsymbol{x}$ are equal to λx^i, it follows, according to (8.68), that $s^2(\lambda \boldsymbol{x}) = \lambda^2 s^2(\boldsymbol{x})$, i.e. $\operatorname{sgn} s^2(\lambda \boldsymbol{x}) = \operatorname{sgn} s^2(\boldsymbol{x})$. It follows from this and from (8.69) that the vector $\lambda \boldsymbol{x}$ is time-like.

We have proved the statement for time-like vectors. Reasoning by analogy, we can make sure that the statement is also true for the case of space-like and isotropic vectors.

The cone of time-like vectors is often designated as T (from the word "time") and the cone of space-like vectors as S (from the word "space").

8.4.2. Galilean coordinates. The Lorentz* transformations. In the theory of pseudo-Euclidean spaces an important role is played by the systems of coordinates in which the *square of the interval* (as it is customary to call the square of the length of the vector $s^2(\boldsymbol{x})$) *has the form*

$$s^2(\boldsymbol{x}) = \sum_{i=1}^{p} (x^i)^2 - \sum_{i=p+1}^{n} (x^i)^2. \tag{8.70}$$

According to the terminology borrowed from physics, *such systems of coordinates are called* ***Galilean***.

The transformations of coordinates which retain expression (8.70) for $s^2(\boldsymbol{x})$ are called the **Lorentz transformations**.

In 8.4.3 we shall discuss the Lorentz transformations of the space $E^4_{(1,3)}$, called the **Minkowski** space.** This space is of interest for physics since it is the space of events of the special theory of relativity.

Note that in a Minkowski space the numeration of the coordinates of a vector usually begins with a zero. Thus, according to (8.70), the square $s^2(\boldsymbol{x})$ of the interval in the space $E^4_{(1,\ 3)}$ is written as follows:

$$s^2(\boldsymbol{x}) = (x^0)^2 - (x^1)^2 - (x^2)^2 - (x^3)^2. \tag{8.71}$$

In physics, for the sake of convenience, the coordinate x^0 is identified with the expression ct, where c is the speed of light and t is a time; x^1, x^2, x^3 are called spatial variables.

In the Minkowski space, the cone T of time-like vectors decomposes into two distinct connected open components T^+ (the cone of the future) and T^- (the cone of the past); the cone S of space-like vectors forms a connected set.

An explanation is due of connected components T^+ and T^-. To do that, we give a physical interpretation of the vector $\boldsymbol{x}$ with the coordinates x^i, $i = 0, 1, 2, 3$ in the space $E^4_{(1,\ 3)}$: this vector is characterized

* Hendrik Antton Lorentz (1853-1928), a Dutch scientist.

** Minkowski, Hermann (1864-1909), a German mathematician and physicist.

by the quantity $\Delta t = x^0/c$ and the vector $\Delta \boldsymbol{r} = \{x^1, x^2, x^3\}$. Thus, regarding $\boldsymbol{x}$ as a displacement in $E^4_{(1,\ 3)}$, we can assume that this displacement is characterized by the displacement Δt in the time and the displacement $\Delta \boldsymbol{r}$ in space.

The time-like vectors $\boldsymbol{x}$ are defined by the condition $s(\boldsymbol{x}) > 0$. In this case, evidently $|\Delta \boldsymbol{r}|/|\Delta t| < c$. If $\Delta t > 0$, then for the displacement of $\boldsymbol{x}$ we get an inequality $0 < |\Delta \boldsymbol{r}|/|\Delta t| < c$. By definition such a displacement of $\boldsymbol{x}$ belongs to T^+ and can be regarded as a displacement of a particle "into the future". If $\Delta t < 0$, then the displacement of $\boldsymbol{x}$ belongs to T^- and can be regarded as a displacement of a particle "into the past" (the motion of anti-particles is explained in this way in physics).

T^+ and T^- are evidently two connected open components of the cone T. The arguments presented explain their names, the cone of the future and the cone of the past.

8.4.3. The Lorentz transformations of the space $E^4_{(1,\ 3)}$. Let us consider a Galilean system of coordinates with the basis $\boldsymbol{e}_i$ in the pseudo-Euclidean space $E^4_{(1,\ 3)}$. In such a system of coordinates, the square of the interval $s^2(\boldsymbol{x})$ has the form (8.71), and the matrix (g_{ij}) of the metric tensor has the form

$$J = (g_{ij}) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}. \tag{8.72}$$

We pass to a new Galilean system of coordinates with the basis $\boldsymbol{e}_{i'}$ and find the conditions which must be satisfied by the coefficients $b^i_{i'}$ of the matrix B of transformation of base vectors. Since

$$\boldsymbol{e}_{i'} = b^i_{i'} \boldsymbol{e}_i, \tag{8.73}$$

and since the matrix $(g_{i'j'})$ of the metric tensor in the basis $\boldsymbol{e}_{i'}$ also has the form (8.72), we can use formulas $g_{i'j'} = b^i_{i'} b^j_{j'} g_{ij}$ for the transformation of the coordinates of a metric tensor and get the following system of equations for determining the coefficients $b^i_{i'}$ of the matrix B of transformation of base vectors (see (8.73)); in this case, the indices i and i' run through the values 0, 1, 2, 3 (see (8.71)):

$$\left.\begin{aligned} &(b^0_{0'})^2 - \sum_{\alpha=1}^{3} b^\alpha_{0'} b^\alpha_{0'} = 1 \\ &b^0_{0'} b^0_{\beta'} - \sum_{\alpha=1}^{3} b^\alpha_{0'} b^\alpha_{\beta'} = 0, \\ &b^0_{\gamma'} b^0_{\beta'} - \sum_{\alpha=1}^{3} b^\alpha_{\gamma'} b^\alpha_{\beta'} = \begin{cases} -1 & \text{for } \gamma' = \beta' \\ 0 & \text{for } \gamma' \neq \beta',\ \gamma',\ \beta' = 1, 2, 3. \end{cases} \end{aligned}\right\} \tag{8.74}$$

We can write relations (8.74) in matrix form. For that purpose we consider, alongside the matrix B, a matrix B^* which can be obtained from B by changing the sign of the elements in the last three columns and a subsequent transposition. Relations (8.74) can evidently be written in the form

$$B^*B = J, \tag{8.75}$$

where the matrix J is defined by relation (8.72) (for the sake of comparison, recall that the matrix C of the orthogonal transformations of a Euclidean space satisfies the relation $C'C = I$, where I is an identity matrix).

Since $\det B^* = -\det B$, and $\det J = -1$, it follows from relation (8.75) that $\det B^*B = \det B^* \det B = -(\det B)^2 = -1$, i.e.

$$\det B = \pm 1. \tag{8.76}$$

We denote by L the collection of all the Lorentz transformations of the Minkowski space. From these general transformations we isolate those which map each vector from T^+ into a vector also belonging to T^+. The collection of such transformations is usually called the **Lorentz transformations of the space** $E^4_{(1,\,3)}$ and is symbolized as $L_\uparrow$.

The general Lorentz transformations for which $\det B = +1$ form a class L_+ of the so-called **proper Lorentz transformations.**

The class L_- of **improper Lorentz transformations** is characterized by the relation $\det B = -1$. The reflection about three spatial axes $x^{1'} = -x^1$, $x^{2'} = -x^2$ and $x^{3'} = -x^3$ can serve as an example of transformations of this kind.

Let B be the matrix of an arbitrary improper Lorentz transformation, and P, the matrix of the reflection just considered. The product of the arbitrary improper transformation and the reflection just considered is, evidently, a proper transformation with the matrix $B' = PB$.

Since $P^2 = P \cdot P = I$, where I is an identity matrix, it follows that

$$B = P^2B = P(PB) = PB'.$$

Thus, *any improper Lorentz transformation is the product of some proper transformation with the matrix B′ and a reflection with the matrix P.*

The intersection of the sets $L_\uparrow$ and L_+ is symbolized as $L_{\uparrow+}$. Some group properties of the sets L, $L_\uparrow$, L_+ and $L_{\uparrow+}$ will be discussed in the next chapter.

In conclusion we seek the transformations $L_{\uparrow+}$ which do not change the coordinates x^2 and x^3. It is clear that these are the transformations $L_\uparrow$ of a two-dimensional pseudo-Euclidean subspace with the coordinates x^0 and x^1, in which the square of the interval can be calculated by the formula $(x^0)^2 - (x^1)^2$.

Let us write formulas (8.74) for the case in question. We get

$$\left.\begin{aligned}&(b^0_{0'})^2-(b'_{0'})^2=1,\\&b^0_{0'}b^0_{1'}-b^1_{0'}b^1_{1'}=0,\\&(b^0_{1'})^2-(b^1_{1'})^2=-1.\end{aligned}\right\}\tag{8.77}$$

Setting $b^1_{0'}/b^0_{0'}=\beta$, we find, from (8.77), the following expressions for the coefficients $b^i_{i'}$ of the transformation matrix B of the basis vectors $\boldsymbol{e}_0$, $\boldsymbol{e}_1$, $\boldsymbol{e}_2$, $\boldsymbol{e}_3$ into the base vectors $\boldsymbol{e}_{0'}$, $\boldsymbol{e}_{1'}$, $\boldsymbol{e}_{2'}$, $\boldsymbol{e}_{3'}$:

$$b^0_{0'}=\pm\frac{1}{\sqrt{1-\beta^2}},\quad b^1_{0'}=\pm\frac{\beta}{\sqrt{1-\beta^2}},$$
$$b^0_{1'}=\pm\frac{\beta}{\sqrt{1-\beta^2}},\quad b^1_{1'}=\pm\frac{1}{\sqrt{1-\beta^2}}.$$

In these formulas the sign is chosen from the condition under which the Lorentz transformation belongs to the class $L_{\uparrow}$. Without going into the details of calculations, we write the final formula for the transformation of coordinates:

$$x^{0'}=\frac{x^0-\beta x^1}{\sqrt{1-\beta^2}},\quad x^{1'}=\frac{x^1-\beta x^0}{\sqrt{1-\beta^2}},\tag{8.78}$$
$$x^{2'}=x^2,\quad x^{3'}=x^3.$$

We set $x^0=ct$, $x^1=x$, $x^2=y$, $x^3=z$, $x^{0'}=ct'$, $x^{1'}=x'$, $x^{2'}=y'$, $x^{3'}=z'$ in (8.78). Then we must rewrite formulas (8.8) as follows:

$$t'=\frac{t-\frac{\beta}{c}x}{\sqrt{1-\beta^2}},\quad x'=\frac{-\beta ct+x}{\sqrt{1-\beta^2}},\quad y'=y,\quad z'=z.\tag{8.79}$$

Let us now elucidate the physical meaning of the constant β.

We assume that the point P is *stationary in the coordinate system* (t', x', y', z'). This means that the time t' varies whereas the spatial coordinates x', y', z' of that point are constant. We shall investigate the behaviour of the point P with respect to the system (t, x, y, z). Differentiating the last three equations (8.79) and taking into account that $dx'=dy'=dz'=0$, we find that $0=\frac{-\beta c\,dt+dx}{\sqrt{1-\beta^2}}$, $0=dy$, $0=dz$. Therefore, $\frac{dx}{dt}=\beta c$, $\frac{dy}{dt}=0$, $\frac{dz}{dt}=0$.

Consequently, *every point P, which is stationary in the coordinate system (t', x', y', z') (and, consequently, the whole system of coordinates), moves relative to the system (t, x, y, z) with a constant velocity $v=\beta c$ in the direction of the Ox axis.*

Thus $\beta = v/c$, where v is the velocity of the movement of the system (t', x', y', z') relative to the system (x, y, z, t). Note that $0 < \beta < 1$ since $0 < v < c$.

Let us rewrite formulas (8.79) as follows:

$$t' = \frac{t - \frac{v}{c^2}x}{\sqrt{1 - \frac{v^2}{c^2}}}, \quad x' = \frac{-vt + x}{\sqrt{1 - \frac{v^2}{c^2}}}, \quad y' = y, \quad z' = z. \tag{8.80}$$

Formulas (8.80) are the transition formulas from the inertial coordinate system (t, x, y, z) to another inertial system (t', x', y', z'). They are known as the **Lorents formulas.**

8.5. The Tensor of the Moment of Inertia

Let us consider a solid fixed at the point O. Suppose $\boldsymbol{r} = \overline{OM}$ is a radius vector of the point M of that solid and $\boldsymbol{v}$ is the velocity of the point M.

As is known, the angular momentum $\boldsymbol{N}$ is defined by the relation $\boldsymbol{N} = \int\limits_V [\boldsymbol{r}\boldsymbol{v}]\, dm$, where V is the volume of the solid and $dm = \rho\, dV$ (ρ is the density of the solid).

Designating the contravariant coordinates of the vector $\boldsymbol{N}$ as N^i and using formula (8.60) for a vector product, we obtain

$$N^i = \int\limits_V c^i_{kl} r^k v^l\, dm \tag{8.81}$$

(recall that $c^i_{kl} = g^{is} c_{skl}$, where c_{skl} are the coordinates of the discriminant tensor in the given basis of the space E^3, see 8.3.6).

By Euler's theorem, there is an instantaneous axis of rotation of a solid. Designating the vector of the instantaneous angular velocity as ω we get $v = [\omega \boldsymbol{r}]$. Using formula (8.60) for a vector product once again, we obtain

$$v^l = c^l_{pn} \omega^p r^n. \tag{8.82}$$

Substituting the expression obtained for v^l into the right-hand side of (8.81) and taking into account that ω^p is independent of the integration variables, we get the following expression for N^i.

$$N^i = \int\limits_V c^i_{kl} c^l_{pn} r^k r^n \omega^p\, dm = \omega^p \int\limits_V c^i_{kl} c^l_{pn} r^k r^n\, dm = \omega^p J^i_p. \tag{8.83}$$

The tensor

$$J^i_p = \int\limits_V c^i_{kl} c^l_{pn} r^k r^n\, dm \tag{8.84}$$

appearing on the right-hand side of (8.83) is known as the **tensor of the moment of inertia.**

Let us transform expression (8.84) for the tensor of the moment of inertia. For that purpose we direct our attention to formula (8.65). According to this formula $c^i_{kl}c^l_{pn}=g_{kn}\delta^i_p-g_{kp}\delta^i_n$. Therefore

$$J^i_p=\int_V (g_{kn}\delta^i_p-g_{kp}\delta^i_n)\,\boldsymbol{r}^n\boldsymbol{r}^k\,dm. \tag{8.85}$$

If we lower the index i in (8.85) with the aid of a metric tensor, we get a frequently used formula for the coordinates of the twice covariant tensor of the moment of inertia:

$$J_{ip}=\int_V (\boldsymbol{r}^2 g_{ip}-\boldsymbol{r}_i\boldsymbol{r}_p)\,dm. \tag{8.86}$$

The tensor of the moment of inertia is widely used in mechanics of a solid body. To give an example, we shall write the expression for the kinetic energy T with the aid of this tensor. We have

$$T=\frac{1}{2}=\int_V v^2\,dm=\frac{1}{2}\int_V g_{ij}v^iv^j\,dm.$$

But sinnce $v^i=c^i_{pn}\omega^p\boldsymbol{r}^n$, the expression for T assumes the form

$$T=\frac{1}{2}\int_V g_{ij}c^i_{pn}c^j_{kl}\omega^p\boldsymbol{r}^n\omega^k\boldsymbol{r}^l\,dm=\frac{1}{2}\,\omega^p\omega^k\int_V g_{ij}c^i_{pn}c^j_{kl}\boldsymbol{r}^n\boldsymbol{r}^l\,dm.$$

From this, in accordance with (8.84), we find that $T=\frac{1}{2}\,\omega^p\omega^kJ_{pk}$.

Chapter 9

ELEMENTS OF THE GROUP THEORY

In this chapter we shall get acquainted with the main concepts of the group theory and study some applications of the theory.

The importance of the group theory is determined by its numerous applications in physics.

9.1. The Concept of a Group. The Main Properties of Groups

9.1.1. Composition laws. We say that a **law of composition** is defined in the set A if we are given a *mapping* T of ordered pairs of elements from A into the set A. In this case, the element c from A, which is put into correspondence with the elements a, b from A with the aid of the mapping $\boldsymbol{T}$, is called the **composition** or **wreath product** of those elements.

The composition c of the elements a and b is symbolized as $a\boldsymbol{T}b$:

$$c = a\boldsymbol{T}b.$$

Other notations are also used for the composition of the elements a and b of the set A. Most frequently used are the *additive* notation $c = a + b$ and the *miltiplicative* notation $c = ab$.

In the case of an additive notation, the corresponding law of composition is called **addition** and in the case of a multiplicative notation it is called **multiplication.** The law of composition is said to be **associative** if for any elements a, b and c of the set A there holds a relation

$$a\boldsymbol{T}(b\boldsymbol{T}c) = (a\boldsymbol{T}b)\boldsymbol{T}c.$$

The composition law is said to be **commutative** if the relation $a\boldsymbol{T}b = b\boldsymbol{T}a$ holds for any pair a, $b \in A$.

The element e of the set A is called **neutral** relative to the law $\boldsymbol{T}$ if the relation $a\boldsymbol{T}e = a$ holds for any element a of the set A.

Ordinary addition and multiplication in the set of real numbers can serve as examples of composition laws. Both these laws are com-

nutative. The neutral element for addition is zero, for multiplication it is unity.

9.1.2. The concept of a group. Some properties of groups. We give the following definition.

Definition 1. *A set A in which the law of composition* $\boldsymbol{T}$ *is defined is called a* ***group*** *G if that law is associative, there is a neutral element e relative to the law* $\boldsymbol{T}$, *and every element a of the set A has an inverse* a^{-1}, *i.e. there is an inverse element such that* $a\boldsymbol{T}a^{-1}=e$.

If we use the multiplicative notation of the composition of elements, Definition 1 assumes the following form.

Definition 2. *A set A of elements a, b, c, ..., in which a law of composition, called multiplication, is defined such that it puts every pair of elements a, b of the set A into correspondence with a definite element* $c=ab$ *of the set, is called a* ***group G*** *if the law satisfies the following requirements:*

1°. $a(bc)=(ab)c$ (*associativity*).

2°. *There is an element e of the set A such that* $ae=a$ *for any element a of that set (the existence of a neutral element).*

3°. *For any element a of the set A there is an inverse element* a^{-1} *such that* $aa^{-1}=e$.

It is customary to call the neutral element E the **identity** element of the group G.

If the composition law $\boldsymbol{T}$ acting in the group G is *commutative*, then the *group G is called* ***commutative*** *or* ***Abelian.*** The additive notation of the composition of elements is most often used for Abelian groups. In that case the neutral element of the Abelian group is called a **zero.**

Let us consider some examples of groups.

(1) The set Z of integers forms an Abelian group with respect to summation.

Indeed, the operation of summation of integers is evidently a law of composition. It is clear that this law is associative and commutative. The integer zero is a neutral element (zero). The integer $-a$ is an inverse element of the integer a.

(2) The set of positive real numbers forms an Abelian group with respect to multiplication. This operation is a law of composition. This law is evidently associative and commutative. The neutral element is the real number unity. The inverse element of the number $a>0$ is $1/a$.

(3) A linear space forms an Abelian group with respect to addition of elements. This operation is a law of composition. According to the axioms of a linear space this law is associative and commutative. The neutral element is a zero element of the space, the inverse element of the element $\boldsymbol{x}$ is $-\boldsymbol{x}$.

(4) Let BCD be an equilateral triangle (see Fig. 9.1). Let us

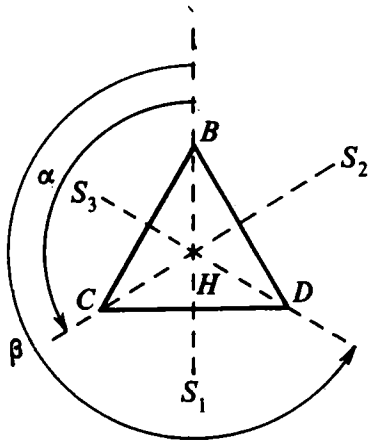


Fig. 9.1

consider the following set A of operations superimposing the triangle on itself:

1°. The rotation α through $2\pi/3$ about the centre H mapping B into C.

2°. The rotation β through $4\pi/3$ mapping B into D.

3°. The symmetry S_1 mapping C into D.

4°. The symmetry S_2 mapping D into B.

5°. The symmetry S_3 mapping B into C.

6°. The identity operation 1.

The following table is a law of composition of the elements of the set A:

	1	α	β	S_1	S_2	S_3
1	1	α	β	S_1	S_2	S_3
α	α	β	1	S_2	S_3	S_1
β	β	1	α	S_3	S_1	S_2
S_1	S_1	S_3	S_2	1	β	α
S_2	S_2	S_1	S_3	α	1	β
S_3	S_3	S_2	S_1	β	α	1

(The rule for using the table can be easily seen from the successive performance of the operations S_1 and then S_2: $S_2S_1=\beta$.)

This composition law is associative but not commutative, there is a neutral element which is an identity operation 1. Every operation has an inverse (there is an identity operation in every row and every column of the table).

Thus the set A of the operations with the indicated law of composition is a group which is evidently noncommutative.

(5) **Permutation groups.**

A one-to-one mapping f of the arbitrary set E onto itself is called a **permutation** of the set E. In the process of permutation, every element a of the set E passes into the element $f(a)$, the inverse permutation f^{-1} maps $f(a)$ into a. The permutation $f(a)=a$ for any a of the set E is called an **identity permutation.**

If the set E consists of elements $a, b, c, \ldots$, then the permutation f of that set is written as follows:

$$f=\begin{pmatrix} a & b & c & \ldots \\ f(a) & f(b) & f(c) & \ldots \end{pmatrix}.$$

The composition law is defined in a natural way in the set P of the permutations of the set E: if f_1 and f_2 are permutations of E, then the successive performance of $f_2 \circ f_1$ number of these permutations is a certain permutation of the set E. It is easy to see that the composition is associative.

If the set P contains an identity permutation, an inverse permutation of every permutation f and, in conjunction with any two permutations f_1, f_2 contains their composition $f_2 \circ f_1$, then, evidently, P is a group.

All the permutations of the set E form a group. For the finite set E consisting of n elements, this group is called a **symmetric group** S_n.

Let us return to example (4) in which we considered the operations of superimposition of the equilateral triangle BCD on itself. We designate the set of the vertices of the triangle as E: $E=(BCD)$.

The group of operations considered in Example (4) can evidently be obtained by the following group of permutations:

$$1=\begin{pmatrix} B & C & D \\ B & C & D \end{pmatrix}, \quad \alpha=\begin{pmatrix} B & C & D \\ C & D & B \end{pmatrix}, \quad \beta=\begin{pmatrix} B & C & D \\ D & B & C \end{pmatrix},$$

$$S_1=\begin{pmatrix} B & C & D \\ B & D & C \end{pmatrix}, \quad S_2=\begin{pmatrix} B & C & D \\ D & C & B \end{pmatrix}, \quad S_3=\begin{pmatrix} B & C & D \\ C & B & D \end{pmatrix}.$$

(6) Let us consider a group Z_2 consisting of two elements 0 and 1 in which multiplication is defined by the rule

$$0\cdot 0=0, \quad 0\cdot 1=1, \quad 1\cdot 0=1, \quad 1\cdot 1=0. \tag{9.1}$$

The unity of the group is the element 0.

This group is known as the **group of residue classes** modulo 2.

(7). Let us consider a group consisting of two elements: (a) an identity transformation of a Euclidean space (we denote this element by 0); (b) the mapping of a Euclidean space with respect to the origin (we denote this element by 1).

The multiplication (i.e. the successive performance of operations (a) and (*b*)) of the elements 0 and 1 is evidently carried out by

rule (9.1). We see that the group in question differs from the group Z_2 (Example 6) only by the nature of the elements. The group properties of these two groups are the same.

Let us note the following *properties of groups* (we use the multiplicative notation for the composition).

Theorem 9.1. *If $aa^{-1}=e$, then $a^{-1}a=e$.*

Proof. Let x be an inverse of a^{-1}:

$$a^{-1}x=e.$$

Then $a=ae=a(a^{-1}x)=(aa^{-1})x=ex$, i.e. $a=ex$. Consequently, $a^{-1}a=a^{-1}(ex)=(a^{-1}e)x=a^{-1}x=e$, i.e. $a^{-1}a=e$. We have proved the theorem.

Theorem 9.2. *The relation $ea=a$ is valid for any element a of the group.*

Proof. By Theorem 9.1, $a^{-1}a=e$ and, in addition, $aa^{-1}=e$. Therefore, $ea=(aa^{-1})a=a(a^{-1}a)=a$, i.e. $ea=a$. We have proved the theorem.

Theorem 9.3. *If $ax=e$ and $ay=e$, then $x=y$.*

Proof. Since $ay=e$, then y is an inverse of a and, therefore, according to Theorem 9.1, $ya=e$. Furthermore, we have $y=ye=y(ax)=(ya)x=ex=x$. We have proved the theorem.

The following important corollaries follow from the theorems we have proved.

Corollary 1. *The inverse of the element a^{-1} is an element a. Or, to put it otherwise, the element a^{-1} is the right as well as the left inverse element of the element a (i.e. $aa^{-1}=e$ and $a^{-1}a=e$).*

Corollary 2. *In any group the equations $ax=b$ and $ya=b$ can be solved uniquely. The solutions of these equations are the elements $x=a^{-1}b$ and $y=ba^{-1}$ respectively.*

Corollary 3. *There is a unique neutral element in a group (the unity of the group) (if $ae=a$ and $ae^*=a$, then $e=e^*$).*

Remark. Note that the *inverse element $(ab)^{-1}$ of the product ab is the element $b^{-1}a^{-1}$.*

Indeed, using the associative property of multiplication, we obtain $(ab)(b^{-1}a^{-1})=a(bb^{-1})a^{-1}=aea^{-1}=aa^{-1}=e$.

9.1.3. Isomorphism of groups. Subgroups. The examples considered in 9.1.2 (see examples 4 and 5, examples 6 and 7) show that there are groups which differ by the nature of their elements but possess the same group properties. Such groups are naturally called **isomorphic**.

We give an exact definition of this concept.

Definition 1. *Two groups G_1 and G_2 are said to be* ***isomorphic*** *if there exists a one-to-one mapping f of the group G_1 onto the group G_2, such that the condition $f(ab)=f(a)f(b)$ holds for any elements a and b of G_1.*

Note that if e_1 is the *unity of group G_1, and e_2 is the unity*

of group G_2, then $f(e_1)=e_2$. Indeed, $f(e_1)=f(e_1e_1)=f(e_1)\cdot f(e_1)$, and the multiplication by the inverse of $f(e_1)$ shows that $e_2=f(e_1)$.

It should also be pointed out that the *inverse mapping* f^{-1} *of group* G_2 *onto group* G_1 *for any elements* x *and* y *of* G_2 *satisfies the condition*

$$f^{-1}(xy)=f^{-1}(x)f^{-1}(y).$$

In addition, it follows from the equation $e_2=f(e_1)=f(aa^{-1})=f(a)f(a^{-1})$ that for any a from G_1 the *inverse of the element* $f(a)$ *is an element* $f(a^{-1})$.

Thus the isomorphic groups considered in abstract form, without indicating the nature of the elements are indistinguished from the viewpoint of group properties.

Remark 1. The correspondence between the isomorphic groups G_1 and G_2 is called an **isomorphism** or **isomorphic mapping** of one group onto the other (the two groups being equivalent).

Remark 2. The isomorphic mapping of group G onto itself is called an **automorphism.**

The automorphisms of a group define in a definite way its symmetry.

If we consider separate automorphisms of a group as some elements and a successive performance of automorphism as the product of the corresponding elements, then the automorphisms themselves form a group (whose unit element is an identity automorphism).

This group is known as the **group of automorphisms** of the given group.

It is easy to ascertain that the group of automorphisms of the group Z_2 (see Example 6 in 9.1.2) is isomorphic to this group.

The concept of a **subgroup** plays an important part in the group theory.

Definition 2. *A subset* G_1 *of elements of the group* G *is called a subgroup of that group if the following conditions are fulfilled:* (1) *if the elements* a *and* b *belong to* G_1, *then* ab *also belongs to* G_1, (2) *if the element* a *belongs to* G_1, *then the inverse element* a^{-1} *also belongs to* G_1.

A subgroup G_1 of the group G, regarded as a set in which the operation of multiplication is defined by the composition law from the enveloping group G, is a group.

It is easy to verify this assertion.

An elementary subgroup of any group is its unit element. The subgroup G_1 of all even numbers in the group G relative to the addition of all integers can serve as another example.

9.1.4. Cosets. Normal subgroups. Let H_1 and H_2 be arbitrary subsets of the group G.

The **product** of the subsets H_1 and H_2 is a subset H_3 consisting of all elements of the form h_1h_2, where $h_1 \in H_1$, $h_2 \in H_2$.

The designation

$$H_3 = H_1 H_2 \tag{9.2}$$

is used for the product of subsets.

Let us consider the case when H_1 consits of a single element h. Then, according to (9.2), the product of H_1 and H_2 can be written as hH_2. Note that if the subsets H_1 and H_2 are subgroups of the group G, then their product H_1H_2 is not, in general, a subgroup.

Suppose H is a subgroup of the group G and a is an element of the group G. The set aH is said to be the **left coset** and the set Ha is said to be the **right coset** of the subgroup H in G.

Of course, when we choose another element instead of a, the right and left cosets of the subgroup H in G change a case.

Note the following properties of cosets (we formulate these properties only for the left cosets, the formulation for the right cosets is similar):

1°. If $a \in H$, then $aH \equiv H$.

2°. The cosets aH and bH coincide if $a^{-1}b \in H$.

3°. Two cosets of the same subgroup H either coincide or do not possess any elements in common.

4°. If aH is a coset, then $a \in aH$.

The first property is obvious. Let us verify the validity of property 2°. Since $a^{-1}bH \equiv H$ according to 1°, we have $bH = (aa^{-1})bH = a(a^{-1}b)H = aH$ because $aa^{-1} = e$. We have thus established property 2°.

Let us prove property 3°. It is, evidently, sufficient to prove that when the cosets aH and bH have an element in common they coincide.

Suppose the elements $h_1 \in H$ and $h_2 \in H$ are such that

$$ah_1 = bh_2 \tag{9.3}$$

(equation (9.3) means that the cosets aH and bH have an element in common). Since H is a subgroup of the group G, the element $h_1h_2^{-1}$ belongs to H. From this and from (9.3) we find that

$$a^{-1}b = h_1h_2^{-1} \in H.$$

Consequently, according to property 2°, $aH = bH$. We have proved property 3°.

Property 4° follows from the fact that the subgroup H contains a unit element e and, therefore, $ae = a \in aH$.

Let H be a subgroup of G for which all the left cosets are simultaneously the right cosets. In this case the relation

$$aH = Ha \tag{9.4}$$

must hold for any element a.

Indeed, according to property 4°, the element $a \in aH$. On the

other hand, the class aH is simultaneously a certain class Hb which evidently (since $a \in Hb$) coincides with the set Ha.

A subgroup H for which all the left cosets are the right-cosets is called a **normal** or **invariant subgroup of group** G.

The following statement holds true.

If H is a normal subgroup of group G, then the product of the cosets is also a coset.

Indeed, let aH and bH be cosets. Then, by the definition of the product of cosets as subsets of the group G, with (9.4) taken into account, we get

$$aHbH = a(Hb)H = a(bH)H = (ab)(HH) = (ab)H,$$

i.e. the product of the cosets $aHbH$ is a coset $(ab)H$.

9.1.5. Homomorphisms. Factor groups.

Suppose G is a group with elements $a, b, c, \ldots$, and $\bar{G}$ is a certain set in which the composition law of its elements $\bar{a}, \bar{b}, \bar{c}, \ldots$ is defined. We shall use the multiplicative notation for the composition: $\bar{a} = \bar{a}\bar{b}$, and the element c is the product of the elements $\bar{a}$ and $\bar{b}$.

Definition 1. *The mapping f of the group G nto the set $\bar{G}$**

$$f : G \to \bar{G} \tag{9.5}$$

*is called a **homomorphic mapping** if the relation*

$$f(ab) = f(a)f(b), \tag{9.6}$$

where $f(a)$, $f(b)$ and $f(ab)$ are the images of the elements a, b and ab under the mapping f, holds for any elements $a \in G$ and $b \in G$.

In this case $\bar{G}$ is a **homomorphic image** or **map** of G.

When $\bar{G}$ is a subset of G, the name **endomorphism** or **endomorphic mapping** is used for homomorphism (9.5).

Remark. If the homomorphic mapping (homomorphism) of the group G onto the set $\bar{G}$ is defined, then all the elements of the group are divided into nonintersecting classes: one class contains all the elements of G which are mapped into the same element of the set $\bar{G}$.

The following statement holds true.

Theorem 9.4. *The homomorphic image of a group is a group.*

Proof. Suppose $\bar{a}, \bar{b}, \bar{c}, \ldots$ are elements of the homomorphic image $\bar{G}$ of the group G under the homomorphic mapping f. This means that there are elements $a, b, c, \ldots$ in the group G such that $\bar{a} = f(a)$, $\bar{b} = f(b)$, $\bar{c} = f(c), \ldots$

* We understand the mapping f of the group G onto the set $\bar{G}$ as the correspondence between the element of the sets G and $\bar{G}$ under which every element $a \in G$ is put into correspondence with only one element $\bar{a} \in \bar{G}$ and every element $a \in \bar{G}$ is an image of at least one element of G. The mapping of G onto $\bar{G}$ is symbolized by means of relation (9.5).

Then multiplication of elements in the set $\bar{G}$ agrees with rule (9.6).

Let us verify that this operation of multiplication meets requirements 1°, 2° and 3° of Definition 2 (see 9.1.2).

1°. *Associativity of multiplication.* Let us form two products $\bar{a}(\bar{b}\bar{c})$ and $(\bar{a}\bar{b})\bar{c}$. We have, according to (9.6),

$$\bar{a}(\bar{b}\bar{c})=f(a)(f(b)f(c))=f(a)f(bc)=f(abc),$$
$$(\bar{a}\bar{b})\bar{c}=(f(a)f(b))f(c)=f(ab)f(c)=f(abc).$$

Comparing these relations, we find that $\bar{a}(\bar{b}\bar{c})=(\bar{a}\bar{b})\bar{c}$. Consequently, the associativity of multiplication of the elements holds true.

2°. *The existence of unity.* We designate as $\bar{e}$ the element $f(e)$, where e is the unity of the group G:

$$\bar{e}=f(e).$$

According to rule (9.6) we have $\bar{a}\bar{e}=f(a)f(e)=f(ae)=f(a)=\bar{a}$ for any element $\bar{a}$ of the set $\bar{G}$.

Consequently, the element $\bar{e}$ indeed plays the part of a unity.

3°. *The existence of an inverse element.* Let us symbolize as $\bar{a}^{-1}$ the element $f(a^{-1})$, where a^{-1} is the inverse of the element a in the group G.

According to (9.6) we have $\bar{a}\bar{a}^{-1}=f(a)f(a^{-1})=f(aa^{-1})=f(e)=\bar{e}$.

Consequently, the element $\bar{a}^{-1}$ plays the part of an inverse of the element $\bar{a}$.

Thus requirements 1°, 2°, 3° of Definition 2 of a group are met for the operation of multiplication of the elements of $\bar{G}$. Therefore, $\bar{G}$ is a group. We have proved the theorem.

Let H be a normal subgroup of the group G. Let us define the following mapping f of the group G onto the set $\bar{G}$ of cosets by the normal subgroup H: if a belongs to G, then we associate this element with the coset to which the indicated element belongs. According to property 3° of cosets (see 9.1.4), under such a mapping every element of the group G is associated with only one class, i.e. we indeed have a mapping f of the group G onto the set of cosets with respect to the normal subgroup H.

Let us prove the following theorem.

Theorem 9.5. *The mapping f of the group G onto cosets by the normal subgroup H indicated above is a homomorphic mapping when the multiplication of cosets is defined as that of subsets of the group G.*

Proof. At the end of 9.1.4 we proved the statement that if aH and bH are cosets, then the product $aHbH$ of those cosets, as subsets of G, is a coset $(ab)H$. Consequently, by means of the mapping f being considered the product ab is put into correspondence

with a coset $(ab)H$ equal to the product of the cosets aH and bH. Therefore, f is a homomorphic mapping. We have proved the theorem.

Corollary. *The set of cosets of the group G by the normal subgroup H with the operation of multiplication of these cosets as subsets of G forms a group.*

This group is called a **factor group** of the group G by the normal subgroup H and is symbolized as G/H.

The validity of the corollary follows from Theorem 9.4.

Remark. The mapping f of the group G onto the set of cosets by the normal subgroup H is evidently a homomorphic mapping of that group onto the factor group G/H.

Let us consider the following **example.**

Let R^n be an n-dimensional linear coordinate space which, as was mentioned in Example 3 of 9.1.2, is an Abelian (i.e. commutative) group with respect to addition of elements (recall that the points x of that space are ordered collections of n real numbers $(x_1, \ldots, x_n)$, the addition of elements $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ being carried out by the rule $(x_1+y_1, \ldots, x_n+y_n)$).

By definition, R^n is a direct product of one-dimensional spaces:

$$R^n = R^1_{(1)} \times R^1_{(2)} \times \ldots \times R^1_{(n)}.$$

Since $R^1_{(n)}$, for example, is an Abelian subgroup, then, evidently, $R^1_{(n)}$ is a normal subgroup of the group R^n. The coset of the element a from R^n is a straight line passing through the point a parallel to the straight line $R^1_{(n)}$, and the factor group $R^n/R^1_{(n)}$ is isomorphic to the $(n-1)$-dimensional subspace R^{n-1}:

$$R^{n-1} = R^1_{(1)} \times R^1_{(2)} \times \ldots \times R^1_{(n-1)}. \tag{9.7}$$

Note that the designation of the factor group $R^n/R^1_{(n)}$ is explained in a certain way by means of the relation

$$R^n/R^1_{(n)} = R^1_{(1)} \times \ldots \times R^1_{(n)}/R^1_{(n)} = R^1_{(1)} \times \ldots \times R^1_{(n-1)}, \tag{9.8}$$

which follows from relation (9.7). It should be pointed out that the last equality sign in formula (9.8) must be regarded as an isomorphism between the respective groups.

We have proved that the homomorphism of the group G onto the factor group G/H is defined by the normal subgroup H. The reverse statement is also true: if the homomorphism of the group G onto the set $\bar{G}$ is defined, then that homomorphism can be used to define the normal subgroup H such that $\bar{G}$* and the factor group G/H are isomorphic.

Let us prove two theorems related to this statement.

---

* According to Theorem 9.4, the homomorphic image of a group is a group.

Theorem 9.6. *Assume that f is a homomorphic mapping of the group G onto $\bar{G}$ and H is a set of elements of the group G which, under the homomorphic mapping f, is mapped into the element $f(e)$, where e is the unity of the group G. Then, H is a normal subgroup of the group G.*

Proof. It is sufficient to prove that H is a subgroup of the group G and every left coset by that subgroup is simultaneously a right coset.

Let us first verify that H is a subgroup of the group G. For that purpose, we must prove that if $a \in H$ and $b \in H$, then $ab \in H$ and also if $a \in H$, then $a^{-1} \in H$.

Assume $a \in H$ and $b \in H$. Since f is a homomorphism we have $f(ab)=f(a)f(b)=f(e)f(e)$.

But $f(e)$ plays the part of a unity in the group $\bar{G}$ (see Theorem 9.4). Therefore, $f(e)f(e)=f(e)$, i.e. $f(ab)=f(e)$. Consequently, $ab \in H$.

Furthermore, assume that $a \in H$, i.e. $f(a)=f(e)$. Then, if a^{-1} is an inverse element of a, then $aa^{-1}=e$, i.e. $aa^{-1} \in H$. Since f is a homomorphism, we have $f(e)=f(aa^{-1})=f(a)f(a^{-1})=f(e)f(a^{-1})=f(a^{-1})$. Therefore, $f(a^{-1})=f(e)$ and, consequently, $a^{-1} \in H$.

Let us prove now that every left coset is at the same time a right coset.

Let a be an arbitrary element of the group G. We shall prove that the set A of the elements of the group G, which are mapped into the element $f(a)$ under the homomorphism f, is simultaneously the left aH and the right Ha coset. And this completes the proof of the theorem.

Assume $a' \in A$. Consider the equation*

$$ax=a'. \tag{9.9}$$

Since f is a homomorphism and $f(a')=f(a)$, we get from this equation $f(ax)=f(a)f(x)=f(a')=f(a)$, i.e. $f(x)=f(e)$. Therefore, $x \in H$. But then, according to (9.9), $a'=ax$, i.e. $a' \in aH$.

Directing our attention to the equation $xa=a'$ and reasoning by analogy, we can verify that $x \in H$. But then $a'=xa$, i.e. $a' \in Ha$. Thus $A=aH=Ha$. We have proved the theorem.

Theorem 9.7 (theorem on the homomorphisms of groups). *Assume that f is a homomorphism of the group G onto $\bar{G}$ and H is the normal subgroup of the group G whose elements correspond to the unity of the group $\bar{G}$* under the homomorphism f. Then the group $\bar{G}$ and the factor group G/H are isomorphic.*

Proof. Let us establish a one-to-one correspondence between the elements of the group $\bar{G}$ and the cosets by the normal subgroup H: we put the element a of the group $\bar{G}$ into correspondence with the

* By virtue of Corollary 2 of Theorem 9.3, this equation is solvable. An element $x=a^{-1}a'$ is its solution.

coset which is mapped into $\bar{a}$ by means of f. This correspondence is evidently one-to-one since, in accordance with property 3° of cosets (see 9.2.4), these cosets do not intersect. If we define the multiplication of these cosets as that of the subgroups of the group $\bar{G}$ and use the assertion proved at the end of 9.1.4, we shall easily see that the one-to-one correspondence we have just established is an isomorphism. But the cosets are exactly elements of a factor group. We have proved the theorem.

9.2. Transformation Groups

We shall study here groups of nonsingular linear transformations of a linear and, in particular, Euclidean space.

9.2.1. Nonsingular linear transformations. The concept of a linear operator was introduced in 5.1.1. Recall that a linear operator $\boldsymbol{A}$ is a mapping of a linear space V into a linear space W under which the image of the sum of elements is equal to the sum of their images and the image of the product of the elements by a number is equal to the product of that number by the image of the element.

We shall consider the so-called ***nonsingular*** *linear operators* mapping the given finite-dimensional linear space V into the same space. The linear operator $\boldsymbol{A}$ is called *nonsingular* if $\det \boldsymbol{A} \neq 0$*.

Note the following *important property of nonsingular operators*: *every such operator maps the space V onto itself one-to-one.*

In other words, *if $\boldsymbol{A}$ is a nonsingular operator, then every element $\boldsymbol{x} \in V$ is associated with only one element $\boldsymbol{y} \in V$ which can be found by the formula*

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}, \tag{9.10}$$

and if $\boldsymbol{y}$ is any fixed element of the space V, then there is only one element $\boldsymbol{x}$ such that $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$.

To prove the second part of the formulated assertion, we shall consider the matrix notation of the action of a linear operator. Thus, if $A = (a_k^j)$ is the matrix of the operator $\boldsymbol{A}$ in a given basis, and the elements $\boldsymbol{x}$ and $\boldsymbol{y}$ have the coordinates $x^1, \ldots, x^n$ and $y^1, \ldots, y^n$ respectively, then, according to formula (5.14) (see 5.2.1), relation (9.10) can be rewritten as

$$y^j = \sum_{k=1}^{n} a_k^j x^k, \quad j = 1, 2, \ldots, n, \tag{9.11}$$

and, therefore, the coordinates x^k can be regarded as the unknowns with the specified coordinates y^j. Since the operator $\boldsymbol{A}$ is nonsingular, i.e. $\det \boldsymbol{A} \neq 0$, system of equations (9.11) has a unique solution for

* Recall that $\det \mathbf{A}$ was introduced in 5.2.2 as the determinant of the matrix of a linear operator in a given basis. It has also been proved there that the value of $\det \boldsymbol{A}$ does not depend on the choice of the basis.

the unknowns x^k. And this means that for every fixed element $\boldsymbol{y} \in V$ there is a unique element $\boldsymbol{x}$ such that $\boldsymbol{y}=\boldsymbol{A}\boldsymbol{x}$.

Thus, the *result of the action of a nonsingular linear operator can be regarded as a mapping of the linear space V onto itself.*

Therefore, *when a nonsingular operator is specified, we can speak of a nonsingular linear transformation of the space V or, simply, of a linear transformation of the space V.*

9.2.2. A group of linear transformations. Assume that V is an n-dimensional linear space with the elements $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$, ... and $GL(n)$ is a set of all nonsingular linear transformations of that space.

Let us determine in $GL(n)$ a composition law which we shall call **multiplication**. We define the multiplication of the linear transformations from $GL(n)$ in the same way as we defined the multiplication of linear operators in 5.1.2.

Namely, the *product* $\boldsymbol{AB}$ *on the linear transformations* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *belonging to the set* $GL(n)$ *is a linear operator acting according to the rule*

$$(\boldsymbol{AB})\boldsymbol{x}=\boldsymbol{A}(\boldsymbol{Bx}). \tag{9.12}$$

Note that in a general case $\boldsymbol{AB}\neq\boldsymbol{BA}$.

For the indicated product to be a composition law indeed (see 9.1.1), it is sufficient to prove that the transformation $\boldsymbol{AB}$ is nonsingular and this follows from the fact that the matrix of the linear operator $\boldsymbol{AB}$ is equal to the product of matrices of the transformations $\boldsymbol{A}$ and $\boldsymbol{B}$ and, consequently, $\det(\boldsymbol{AB})=\det\boldsymbol{A}\cdot\det\boldsymbol{B}\neq 0$ since $\det\boldsymbol{A}\neq 0$ and $\det\boldsymbol{B}\neq 0$.

Let us prove now the following theorem.

Theorem 9.8. *The set* $GL(n)$ *of nonsingular linear transformations of the linear n-dimensional space V with the operation of multiplication introduced above is a group (called a* ***group of linear transformations*** *of the linear space V).*

Proof. Let us verify requirements 1°, 2°, 3° of Definition 2 of a group (see 9.1.2).

1°. *Associativity of multiplication*, i.e. the equation

$$\boldsymbol{A}(\boldsymbol{BC})=(\boldsymbol{AB})\boldsymbol{C}$$

is valid since, according to (9.12), the product of linear transformations consists in their successive actions and, therefore, the linear transformations $\boldsymbol{A}(\boldsymbol{BC})$ and $(\boldsymbol{AB})\boldsymbol{C}$ coincide with the linear transformation $\boldsymbol{ABC}$ and are consequently identical.

2°. *The existence of a unity.* We denote an identity transformation by $\boldsymbol{I}$. This transformation is nonsingular since $\det\boldsymbol{I}=1$. The equality $\boldsymbol{AI}=\boldsymbol{IA}=\boldsymbol{A}$ is evidently valid for any transformation $\boldsymbol{A}$ from $GL(n)$.

Consequently, the linear transformation $\boldsymbol{I}$ plays the part of a unity.

3°. *The existence of an inverse element.* Assume that $\boldsymbol{A}$ is any

fixed nonsingular linear transformation. Let us direct our attention to coordinate notation (9.11) of that transformation. Since $\det \boldsymbol{A} \neq 0$, we can use system (9.11) to define uniquely the determinant $\boldsymbol{x}$ (coordinates x^k) from the given $\boldsymbol{y}$ (from the given coordinates y^j). Consequently, we have defined the inverse transformation $\boldsymbol{A}^{-1}$ which is evidently linear (this follows from (9.11)); in addition $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$ by the definition itself. Therefore, the linear operator $\boldsymbol{A}^{-1}$ plays the part of the inverse operator for $\boldsymbol{A}$.

Thus requirements 1°, 2°, 3° from Definition 2 of a group are met for the operation of multiplication of elements from $GL(n)$. Therefore, $GL(n)$ is a group. We have proved the theorem.

9.2.3. Convergence of elements in the group $GL(n)$. The subgroups of the group $GL(n)$. In this and the following subsections we shall consider the group $GL(n)$ in an n-dimensional Euclidean space V.

We introduce the concept of convergence in the group $GL(n)$.

Definition. *The sequence of the elements $\{\boldsymbol{A}_n\}$ from $GL(n)$ is said to be* ***convergent*** *to the element $\boldsymbol{A} \in GL(n)$ if the sequence $\{\boldsymbol{A}_n\boldsymbol{x}\}$ converges to $\boldsymbol{A}\boldsymbol{x}$ for any $\boldsymbol{x}$*.*

We shall use the concept of convergence in $GL(n)$ later on when we introduce the so-called *compact* groups.

The subgroups of $GL(n)$ are classified as follows.

1°. *Finite subgroups*, i.e. subgroups containing a finite number of elements.

A subgroup of reflections about the origin containing two elements, an identity transformation and a reflection about the origin (see Example 7 in 9.1.2) can serve as an example of a finite subgroup.

2°. *Discrete subgroups*, i.e. subgroups containing denumerably many elements.

A subgroup of rotations of a plane about the origin through the angles $k\varphi$, $k = 0, \pm 1, \pm 2, \ldots$, where φ is an angle incommensurable with π, can serve as an example of such a subgroup.

3°. *Continuous subgroups*, i.e. subgroups containing more than a denumerable number of elements.

The subgroup of all rotations of a three-dimensional space about a fixed axis is an example of a continuous subgroup.

From the continuous subgroups of the group $GL(n)$ we can isolate the so-called **compact subgroups**, i.e. subgroups in which we can isolate, from any infinite set of elements, a sequence converging to an element of that subgroup.

9.2.4. A group of orthogonal transformations. In the group

* The sequence $\{\boldsymbol{A}_n\boldsymbol{x}\}$ is a sequence of points of the space V. The convergence of the sequence $\{\boldsymbol{A}_n\boldsymbol{x}\}$ is, therefore, understood in the ordinary sense.

$GL(n)$ we can isolate a special subgroup of the so-called **orthogonal transformations.** These transformations, regarded as a separate set, form a group called an orthogonal group.

We introduce the concept of orthogonal transformations.

Recall that we consider nonsingular linear transformations. The concept of such a transformation is equivalent to the notion of a nonsingular operator, i.e. the operator $\boldsymbol{A}$ for which $\det \boldsymbol{A} \neq 0$.

Recall now the concept of an orthogonal operator acting in the real Euclidean space V introduced in 5.9.

Namely, we called the linear operator $\boldsymbol{P}$ orthogonal if the relation

$$(\boldsymbol{P}\boldsymbol{x},\ \boldsymbol{P}\boldsymbol{y}) = (\boldsymbol{x},\ \boldsymbol{y}) \tag{9.13}$$

held for any $\boldsymbol{x}$ and $\boldsymbol{y}$.

The result of the action of the orthogonal operator $\boldsymbol{P}$ is an orthogonal transformation $\boldsymbol{P}$.

It was proved in Theorem 5.36 that the operator $\boldsymbol{P}$ was orthogonal if and only if there was an inverse operator $\boldsymbol{P}^{-1}$ and there held an equality

$$\boldsymbol{P}^{-1} = \boldsymbol{P}^* \tag{9.14}$$

In this equation $\boldsymbol{P}^*$ is an adjoint operator of $\boldsymbol{P}$.

Thus, if the transformation $\boldsymbol{P}$ is orthogonal, then it has an inverse $\boldsymbol{P}^{-1}$. Hence it follows that *every orthogonal transformation is nonsingular.* Indeed, since $\boldsymbol{P}\boldsymbol{P}^{-1} = \boldsymbol{I}$, where $\boldsymbol{I}$ is an identity transformation, we have

$$\det \boldsymbol{P} \cdot \det \boldsymbol{P}^{-1} = \det \boldsymbol{I} = 1,$$

i.e. $\det \boldsymbol{P} \neq 0$. Consequently, the orthogonal transformation $\boldsymbol{P}$ is nonsingular.

Note the following important property of orthogonal transformations.

Theorem 9.9. *The set of all orthogonal transformations of the Euclidean space V with an ordinary operation of multiplication of linear transformations forms a group (called an* ***orthogonal*** *group and symbolized as $O(n)$).*

Proof. It is sufficient to prove that the product of orthogonal transformations is an orthogonal transformation. The existence of an inverse transformation (an inverse element) for a given orthogonal transformation was proved in Theorem 5.36.

Thus, let $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ be orthogonal transformations. Let us consider the product $\boldsymbol{P}_1\boldsymbol{P}_2$. According to Theorem 5.36, we have only to prove the relation

$$(\boldsymbol{P}_1\boldsymbol{P}_2)(\boldsymbol{P}_1\boldsymbol{P}_2)^* = \boldsymbol{I}. \tag{9.15}$$

We established in 5.5.1 (see property 5° of adjoint operators) that $(\boldsymbol{P}_1\boldsymbol{P}_2)^* = \boldsymbol{P}_2^*\boldsymbol{P}_1^*$. Using this relation and the orthogonality of the

transformations $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$, we get

$$(\boldsymbol{P}_1\boldsymbol{P}_2)(\boldsymbol{P}_1\boldsymbol{P}_2)^*=(\boldsymbol{P}_1\boldsymbol{P}_2)(\boldsymbol{P}_2^*\boldsymbol{P}_1^*)=\boldsymbol{P}_1(\boldsymbol{P}_2\boldsymbol{P}_2^*)\boldsymbol{P}_1^*=\boldsymbol{P}_1\boldsymbol{I}\boldsymbol{P}_1^*=\boldsymbol{P}_1\boldsymbol{P}_1^*=\boldsymbol{I}.$$

We have proved relation (9.15) and thus proved the theorem.

Remark 1. The orthogonal group is evidently a subgroup of the group $GL(n)$.

Remark 2. The value of the determinant $\det \boldsymbol{P}$ of the orthogonal transformation $\boldsymbol{P}$ satisfies the relation

$$(\det \boldsymbol{P})^2=1. \tag{9.16}$$

Thus,

$$\det \boldsymbol{P}=\pm 1. \tag{9.17}$$

To prove (9.16), we must keep in mind that the relation

$$PP'=I, \tag{9.18}$$

where P' is a transposed matrix obtained from P by an interchange of rows and columns and I is an identity matrix, holds for the transformation $\boldsymbol{P}$.

Since $\det P=\det P'$ (a determinant does not change upon an interchange of rows and columns) and $\det I=1$, it follows from relation (9.18) that $(\det P)^2=1$, i.e. $\det P=\pm 1$. Since by definition $\det \boldsymbol{P}$ is introduced as the determinant of the matrix P in any basis, we have proved relations (9.16) and (9.17).

Relation (9.17) for the determinant of an orthogonal transformation serves as the basis for the division of all these transformations into two classes.

In the first class we place all orthogonal transformations for which $\det \boldsymbol{P}=+1$. In what follows, we shall call these transformations **proper.**

In the second class we place all orthogonal transformations for which $\det \boldsymbol{P}=-1$. These transformations are called **improper.**

The set of all proper orthogonal transformations forms a group called a **special** or **proper orthogonal group.** It is symbolized as $SO(n)$.

We can prove that every group $SO(n)$ is compact.

9.2.5. Certain discrete and finite subgroups of an orthogonal group. We shall not try to make the exposition here as complete as possible. We shall cite examples showing the characteristics of some subgroups of an orthogonal group.

Note that finite and discrete subgroups of the group $O(3)$ find important application in crystallography.

1°. Let us consider a two-dimensional orthogonal group $O(2)$. From this group we can isolate a discrete group of rotations through the angle $k\varphi$, $k=0, \pm 1, \pm 2, \ldots$

We designate as a the element of this subgroup corresponding to the value $k=\pm 1$. Then the element a_k, corresponding to the rotation through the angle $k\varphi$ for $k>0$, is evidently equal to $a_k=\underbrace{a\cdot a\ldots a}_{k \text{ times}}$. This relation can be abbreviated as

$$a_k=a^k,\ k=1,\ 2,\ \ldots$$

If we symbolize as a^{-1} the inverse of the element a (a^{-1} is the element corresponding to the rotation through the angle $-\varphi$) and designate the unity of the subgroup being considered as a^0, then, evidently, any element a_k for a positive, negative and zero value of k can be written as

$$a_k=a^k,\ k=0,\ \pm 1,\ \pm 2,\ \ldots \tag{9.19}$$

The groups whose elements a_k can be represented in the form (9.19) are called **cyclic groups.**

Cyclic groups are evidently *discrete.*

Note the following two types of the cyclic subgroups of rotation.

(1) If $\varphi\neq 2\pi p/q$, where p and q are integers (i.e. the angle is incommensurable with π), then all the elements a_k are distinct.

(2) If $\varphi=2\pi p/q$, where p and q are coprime numbers, then the relation $a_{k+q}=a_k$ holds true, i.e. $a^q=a^0$.

The groups for which the last relation holds true are called the **cyclic groups of order q.**

2°. Let us direct our attention now to the so-called subgroups of **mirror symmetry.**

Every subgroup of mirror symmetry consists of two elements: a unity (an identity transformation) and a reflection about some plane or about the origin.

It is relatively easy to verify that an identity transformation and a reflection form a group. To do that, it is sufficient to note that two successive reflections produce an identity transformation (see Example 7 in 9.1.2).

Let us consider, for instance, a subgroup $\{\boldsymbol{I},\ \boldsymbol{P}\}$ of the group $O(3)$ consisting of a unity $\boldsymbol{I}$ and a reflection $\boldsymbol{P}$ of a three-dimensional space about the origin. In an orthonormal basis the matrix P of that transformation has the form $P=\begin{pmatrix}-1 & 0 & 0\\ 0 & -1 & 0\\ 0 & 0 & -1\end{pmatrix}$.

Since the determinant $\det \boldsymbol{P}=-1$, the subgroup $\{\boldsymbol{I},\ \boldsymbol{P}\}$ is improper. We pointed out in Example 7 in 9.1.2 that the subgroup $\{\boldsymbol{I},\ \boldsymbol{P}\}$ was isomorphic to the group Z_2 of residues modulo 2.

Let us prove the following statement.

The subgroup $\{\boldsymbol{I},\ \boldsymbol{P}\}$ being considered is a normal subgroup of the group $O(3)$.

We have to prove that the relations

$$aI = Ia, \quad aP = Pa \tag{9.20}$$

hold for any element a from $O(3)$ (these relations show that the left and right cosets of the subgroup $\{I, P\}$ coincide, which is an indication of a normal subgroup).

The first relation (9.20) is obvious.

To prove the second relation (9.20), we make use of the following obvious properties of the reflection P:

$$PP = I, \quad PaP = a, \quad a \in O(3).$$

Multiplying the relation PaP on the left by P and using equation $PP = I$, we get the second relation (9.20).

Let us now prove the following assertion.

The subgroup $SO(3)$ of proper orthogonal transformations of the group $O(3)$ is isomorphic to the factor group $O(3)$ by the normal subgroup $\{I, P\}$.

Proof. The coset of the element $a \in SO(3)$ by the subgroup $\{I, P\}$ has the form $\{a, Pa\}$, Pa being an improper transformation (the product of the proper transformation a by the improper transformation P is an improper transformation).

If a' is an improper transformation, then the coset $\{a', Pa'\}$ can be reduced to the form $\{a, Pa\}$, where $a = Pa'$ is a proper transformation and $Pa = P(Pa') = (PP)a' = a'$.

Thus the factor group $O(3)/\{I, P\}$ consists of cosets of the form $\{a, Pa\}$, where a is a proper transformation. The relation $a \leftrightarrow \{a, Pa\}$ is evidently an isomorphism between the groups $SO(3)$ and $O(3)/\{I, P\}$. We have proved the assertion.

9.2.6. The Lorentz group. We introduced in 8.4.1 the concept of a pseudo-Euclidean space $E^n_{(p,q)}$, i.e. a linear space in which a scalar product (x, y) equal to the nonsingular symmetric bilinear form $A(x, y)$, the polar of the alternating quadratic form $A(x, x)$, is specified:

$$(x, y) = A(x, y). \tag{9.21}$$

We pointed out in 8.4.2 that in the so-called Galilean system of coordinates the square of the interval

$$s^2(x) = (x, x) \tag{9.22}$$

(it is an accepted term for the square of the length of the vector x with the coordinates $(x_1, x_2, \ldots, x_n)$) had the form

$$s^2(x) = \sum_{i=1}^{p} x_i^2 - \sum_{i=p+1}^{n} x_i^2. \tag{9.23}$$

We must introduce the concept of the **Lorentz transformation** of the pseudo-Euclidean space $E^n_{(p,q)}$.

Definition. *A linear transformation* $\boldsymbol{P}$ *of the pseudo-Euclidean space* $E^n_{(p,q)}$ *is called a* ***Lorentz transformation*** *if for any* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *from* $E^n_{(p,q)}$ *there holds a relation*

$$(\boldsymbol{Px},\ \boldsymbol{Py})=(\boldsymbol{x},\ \boldsymbol{y}), \tag{9.24}$$

where $(\boldsymbol{x},\ \boldsymbol{y})$ *is a scalar product defined by relation* (9.21).

Equation (9.24) is a **condition for a Lorentz transformation.**

Note that under the Lorentz transformation the square of the interval $s^2(\boldsymbol{x})$, defined by relations (9.22) (or (9.23)) is retained.

As in 9.2.4, we can prove that the determinant $\det \boldsymbol{P}$ of a Lorentz transformation is nonzero and, therefore, for every Lorentz transformation there is an inverse transformation $\boldsymbol{P}^{-1}$.

In addition, by the meaning of the definition of a Lorentz transformation the product of such transformations is a Lorentz transformation. Thus the following assertion holds true.

The set of all Lorentz transformations of the pseudo-Euclidean space $E^n_{(p,q)}$ *with an ordinary operation of multiplication of linear transformations (linear operators) forms a group which is called a* ***general Lorentz group*** *of the pseudo-Euclidean space* $E^n_{(p,q)}$ *and is designated as* $L(n,\ p,\ q)$.

We isolate a special class of pseudo-Euclidean spaces $E^n_{(1,\,n-1)}$ (it includes a space $E^4_{(1,3)}$ which is interesting from the point of view of physics).

The Lorentz group for the spaces $E^n_{(1,\,n-1)}$ is designated as $L(n)$.

In 8.4.1 (formula (8.69)) we introduced the concept of the length $\sigma(\boldsymbol{x})$ of the vector $\boldsymbol{x}$, which can be calculated by the formula

$$\sigma(\boldsymbol{x})=(\operatorname{sgn} s^2(\boldsymbol{x}))\sqrt{|\,s^2(\boldsymbol{x})\,|}\,.$$

By means of this formula all nonzero vectors of a pseudo-Euclidean space are divided into **time-like** ($\sigma(\boldsymbol{x})>0$), **space-like** ($\sigma(\boldsymbol{x})<0$) and **isotropic** ($\sigma(\boldsymbol{x})=0$). It was proved that the set of the termini of the time-like (space-like, isotropic) vectors whose origins coincide with an arbitrary fixed point, forms a cone T (the cone of space-like vectors is denoted by S). By agreement*, the cone T is divided into two connected components T^+ and T^- (the cone of the future and the cone of the past) so that each component, together with the vector $\boldsymbol{x}$, contains any vector $\lambda\boldsymbol{x}$, where $\lambda>0$.

* It is indicated in 8.4.2. for the Minkowski space $E^4_{(1,\,3)}$ how the cone T is divided into the connected components T^+ and T^- and the physical interpretation of those components is given.

The division of vectors in a pseudo-Euclidean space just described makes it possible to separate some subgroups from the Lorentz group $L(n)$.

Namely, the subgroup $L(n)$ whose transformations carry any time-like vector again into a time-like vector is called a **complete Lorentz group.** It is designated as $L_{\uparrow}(n)$.

Another subgroup of the group $L(n)$ can be distinguished. It contains transformations the determinant of whose matrix is positive. This subgroup is designated as $L_{+}(n)$ and is called a **proper Lorentz group.**

The proper Lorentz transformations which belong to the subgroup $L_{\uparrow}(n)$ also form a subgroup. It is often called a **Lorentz group** and designated as $L_{\uparrow}(n)$.

To complete this subsection, we want to point out that *as distinct from orthogonal groups the Lorentz groups are noncompact**.

We shall prove, for example, the noncompactness of the group $L_{\uparrow}(2)$.

We completely described this group in 8.4.3. Recall that if a system of coordinates (x, y) has been introduced in $E^2_{(1,1)}$ such that the square of the interval is specified by the formula

$$s^2 = x^2 - y^2, \tag{9.25}$$

then the Lorentz transformations from the group $L_{\uparrow}(2)$ of the space $E^2_{(1,1)}$ are specified by the formulas

$$x' = \frac{x - \beta y}{\sqrt{1-\beta^2}}, \quad y' = \frac{y - \beta x}{\sqrt{1-\beta^2}}, \tag{9.26}$$

Let us consider a vector $\boldsymbol{x}$ with coordinates $(0, 1)$ in the plane (x, y). By formula (9.26), this vector passes into a vector $\boldsymbol{x}_\beta$ with coordinates

$$\left(\frac{-\beta}{\sqrt{1-\beta^2}}, \frac{1}{\sqrt{1-\beta^2}}\right). \tag{9.27}$$

---

* In 9.2.3 we introduced the concept of convergence of elements into the group $GL(n)$ in an n-dimensional Euclidean space and the related concept of a compact group.

These concepts can be easily extended to the case of a group in an arbitrary finite-dimensional linear space V. We first introduce the concept of convergence of points in V (say, we can choose in V a system of coordinates and consider the convergence of the sequence of vectors $\{\boldsymbol{x}_m\}$ as the convergence of the sequences of the coordinates of those vectors). After that, by a complete analogy with the definition of convergence in the case of the group $GL(n)$, in a Euclidean space, we introduce the concept of convergence in a group given in a linear space and define the concept of a compact

Let us consider now the sequence of the Lorentz transformations (9.26) defined by the values β_n from the relation

$$1-\beta_n^2=\frac{1}{n}, \quad n=1, 2, \ldots \tag{9.28}$$

In accordance with (9.27) and (9.28), under the action of the indicated sequence of the Lorentz transformations, the vector $\boldsymbol{x}$ passes into the following sequence of the vectors $\{\boldsymbol{x}_n\}$ with coordinates

$$(-\sqrt{n-1}, \sqrt{n}). \tag{9.29}$$

Thus, for the values β from equation (9.28) we cannot isolate a convergent sequence from an infinite sequence of the Lorentz transformations in the group $L_{\uparrow}(2)$ defined by relations (9.26) (recall that the sequence of the elements $\boldsymbol{A}_n$ of a group is said to be convergent to the element $\boldsymbol{A}$ if the sequence $\{\boldsymbol{A}_n\boldsymbol{x}\}$ converges to $\boldsymbol{A}\boldsymbol{x}$ for any $\boldsymbol{x}$) since sequence (9.29) is unbounded.

The geometrical illustration of the noncompactness of the group $L_{\uparrow}(2)$ consists in the following.

In accordance with (9.25), a circle of radius unity in a pseudo-Euclidean plane is a hyperbola $x^2-y^2=1$ which is a noncompact set. After a sequence of transformations from the group $L_{\uparrow}(2)$ considered above, the given point on this circle is transformed into an infinitely large sequence of points on the hyperbola indicated above, and we cannot isolate a convergent sequence from an infinitely large sequence of points.

9.2.7. Unitary groups. We shall consider here a complex linear space. By a complete analogy with 9.2.2, we can consider groups of linear transformations of that space. Since a complex number is defined by two real numbers (the real and the imaginary part), the complete linear group $GL(n)$ of transformation of an n-dimensional complex linear space is isomorphic to the complete linear group of transformations of the real $2n$-dimensional space $GL(2n)$ (we can replace this symbol by $GL(2n, R)$ emphasizing that we speak of a group of transformations of a real space).

By analogy with a real Euclidean space, we regard, in a complete linear group of transformations of a complex Euclidean space, the so-called unitary groups $U(n)$ which are an analog of orthogonal groups (recall that in 5.7 unitary transformations (unitary operators) were defined as linear transformations retaining a scalar product; orthogonal transformations in a real case were defined in the same way).

As in a real case, we can isolate in the group $U(n)$ of unitary transformations a subgroup $SU(n)$ for which the determinants of unitary transformations are equal to unity.

9.3. Group Representations

In 9.2 we considered groups of linear transformations of a linear space. Thus we investigated linear transformations from the point of view of their group properties. We have also considered geometrical and other properties of linear transformations.

We shall be interested here, in a definite sense, in a converse question: to what extent the properties of an abstractly defined group can be characterized by means of groups of linear transformations.

One of the means of solving this problem consists in the homomorphic (and, in particular, an isomorphic) mapping of an abstract group onto a subgroup (or the entire group) of linear transformations.

Thus we arrive at the concept of **representation** of a given group by means of a subgroup of the group of linear transformations*.

The study of various representations of a given group makes it possible to elucidate significant properties of the group which are important for applications.

In many branches of physics (crystallography, the theory of relativity, quantum mechanics, etc) it is sometimes necessary to construct representations of various groups (finite and discrete subgroups of the group $GL(3)$, groups $O(3)$, $L(3)$, $U(3)$, $SU(3)$, etc). These constructions are far beyond the scope of the fundamental concepts of the group theory and cannot be discussed in this book.

We shall restrict our consideration to some concepts used in the theory of representations and to some examples.

9.3.1. Linear representations of groups. Terminology.

Definition. *A **linear representation** of the group G in a finite-dimensional Euclidean space E^n is a mapping f which puts every element a of that group into correspondence with the linear transformation T_a of the space E^n such that the relation $T_{(a_1a_2)}=T_{a_1}T_{a_2}$ holds for any a_1 and a_2 belonging to G.*

Thus the linear representation of the group G in the finite-dimensional Euclidean space E^n is a homomorphic mapping of that group onto a certain subset of the linear transformations of that space.

The following terminology is used: the space E^n is called the **space of representation**, the dimension n of that space is called the **dimension of the representation**, and the basis in the space E^n is called the **basis of the representation**.

Note that the homomorphic image $f(G)$ of the group G is also

* We can also consider a more general problem of representation of a given group by mapping it onto some group of transformations.

called the representation of that group in the space of representations.

In what follows, for the sake of brevity, n-dimensional linear representations of a group will be simply called the representations of that group.

The symbol $D(G)$ is used to designate the representation of the group G; various representations of the given group are denoted by indices (say, $D^{(\mu)}(G)$). The symbol $D^{(\mu)}(g)$ is used for the linear transformation (linear operator) corresponding to the element $g \in G$ in the representation $D^{(\mu)}(G)$.

The **trivial representation of the group G** is a homomorphic mapping of G into a unit element of the group $GL(n)$.

If the mapping f of the group G onto the subgroup $GL(n)$ is isomorphic, the representation is called **exact.**

It is evident that not every group has an exact n-dimensional representation for a given n. For example, the group $O(10)$ cannot have an exact one-dimensional representation (this follows, in particular, from the fact that the group $O(1)$ is Abelian and the group $O(10)$ is not Abelian).

Note that the representation of a group resulting from the homomorphic mapping f of the group G into $GL(n)$ is isomorphic to the factor group $G/\text{kern } f$, where kern f is the so-called congruence kernel of the homomorphism f, i.e. the set of the elements of G which is mapped into the unity of the group $GL(n)$ under the homomorphic mapping f.

9.3.2. Matrices of linear representations. Equivalent representations. Let us consider the representation $D^{(\mu)}(G)$ of the group G. In this representation, every element g from G is associated with the linear transformation $D^{(\mu)}(g)$. We designate the matrix of this linear transformation in the basis of the representation $D^{(\mu)}(G)$ as $D_j^{(\mu)i}(G)$ or $D_{ij}^{(\mu)}(g)$.

The matrix $D_{ij}^{(\mu)}(g)$ corresponding to the element g changes with the choice of the basis in the space of the representations. The question concerning the **equivalence of representations** of the group in the same space naturally arises.

Here is the definition of the equivalence of representations.

Definition. *The representations $D^{(\mu_1)}(G)$ and $D^{(\mu_2)}(G)$ of the group G in the same space E^n are said to be* ***equivalent*** *if there is a non-singular linear transformation C of the space E^n such that the relation $D^{(\mu_1)}(g) = C^{-1}D^{(\mu_2)}(g)C$ holds for every element $g \in G$.*

The concept of equivalence plays a significant role in the theory of representations, mainly in the enumeration and classification of representations.

The choice of a basis in the space of representations is also significant because in some basis the matrices corresponding

to the elements of the group may have a standard, sufficiently simple form, which enables us to make important inferences concerning the representation under investigation. In 9.3.3 we shall give a certain classification of representations proceeding from a special form of matrices.

9.3.3. Reducible and irreducible representations. We shall discuss here the conditions under which the given representation $D(G)$, defined in the space E^n, induces a representation $\tilde{D}(G)$ in the subspace E' of that space.

This problem is closely connected with that of describing a given representation by means of more simple representations which have a smaller dimension than the given representation.

The concept of an invariant subspace of a linear transformation (linear operator) is closely connected with the solution of this problem.

Recall that the subspace E' is an invariant subspace of the linear operator $\boldsymbol{A}$ if for every element $\boldsymbol{x}$ from E' the element $\boldsymbol{Ax}$ belongs to E' (see 5.3). In other words, the subspace E' is invariant if the action of the operator $\boldsymbol{A}$ on the elements of that subspace does not take them outside that subspace. Note that the space E^n itself and the zero element of the space are invariant subspaces of any linear operator.

We can introduce the concept of an invariant subspace for the representation $D(G)$. Namely, *the subspace E' is said to be* ***invariant*** *for the representation $D(G)$ if it is invariant for every operator from $D(G)$.*

It is evident that a certain representation $\tilde{D}(G)$ is induced on the invariant subspace of the representation $D(G)$. It should be pointed out that the representation $\tilde{D}(G)$ does not reduce to the representation $D(G)$ if the invariant subspace E' does not coincide with E^n.

Let us discuss the concept of a **reducible** representation in more detail.

Suppose, for instance, that all the matrices of a certain three-dimensional representation $D(G)$ have the form

$$\left(\begin{array}{c|c} A_1 & A_2 \\ \hline 0 & A_3 \end{array}\right) = \left(\begin{array}{cc|c} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ \hline 0 & 0 & a_{33} \end{array}\right), \tag{9.30}$$

where A_1, A_2, A_3 and O denote the matrices $\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$, $\begin{pmatrix} a_{13} \\ a_{23} \end{pmatrix}$, (a_{33}) and $(0, 0)$ respectively. It is easy to verify that the product of the matrices of form (9.30) obeys the law

$$\left(\begin{array}{c|c} A_1' & A_2' \\ \hline 0 & A_3' \end{array}\right)\left(\begin{array}{c|c} A_1'' & A_2'' \\ \hline 0 & A_3'' \end{array}\right) = \left(\begin{array}{c|c} A_1'A_1'' & A_2''' \\ \hline 0 & A_3'A_3'' \end{array}\right),$$

i.e. the product of the matrices of form (9.30) is a matrix of the form (9.30). Moreover, when matrices of this form are multiplied, the matrices A_1' and A_1'' and the matrices A_3' and A_3'' are multiplied between themselves.

Thus we see that the matrix $A_1 = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ forms a two-dimensional representation of the group being considered, and the matrix $A_3 = (a_{33})$ forms a one-dimensional representation of the same group.

In these cases we say that the representation $D(G)$ is **reducible.**

If all the matrices (we speak of square matrices of order n) of the operators of the representation have the form

$$\left(\begin{array}{c|c} A_1 & 0 \\ \hline 0 & A_2 \end{array}\right), \tag{9.31}$$

where A_1 and A_2 are square matrices of different orders in a general case, then it is clear that the matrices A_1 and A_2 form representations the sum of whose dimensions is equal to n.

In this case the representation is **fully reducible**. Note that the operators whose matrices have the form (9.31) are actually reduced to two operators acting independently in two invariant subspaces.

It should also be pointed out that the representation which is induced on an invariant subspace by the given representation $D(G)$ is called a **part** of the representation $D(G)$.

In conclusion, we shall formulate the concept of an **irreducible** representation.

The representation $D(G)$ of the group G is said to be ***irreducible*** *if it has only two invariant subspaces,* E^n and O.

Otherwise, the representation is **reducible.**

The part played by irreducible representations consists in the fact that any representation can be expressed in terms of irreducible representations.

9.3.4. Characters. In the theory of representations of groups and, in particular, in the theory of representations of finite groups a useful role is played by the invariants of linear transformations which form a representation. The importance of the invariants is also clear from the fact that they do not depend on the choice of the basis of a representation and, therefore, they characterize the representation in a certain sense.

Assume that $D(G)$ is an n-dimensional representation of the

group G, and $D^i_j(g)$ is the matrix of the operator corresponding to the element g from G.

The ***character*** *of the element $g \in G$ in the representation $D(G)$ is the number*

$$\chi(g)=D^i_i(g)=D^1_1(g)+D^2_2(g)+\ldots+D^n_n(g).$$

Thus *the character of the element g is the trace of the matrix of the operator $D(g)$.*

Since the trace of the matrix of a linear operator is an invariant (see 5.2.3), *the character of any element does not depend on the basis of the representation and is, therefore, invariant.*

Thus every element $g \in G$ of the representation $D(G)$ is associated with a number which is the character of that element.

Since different elements may have the same characters, we must find which elements of the group are associated with the same characters. To solve this problem, we introduce the concept of conjugate elements and the classes of conjugate elements in the given group G.

The element $b \in G$ is said to be the ***conjugate*** *element of $a \in G$ if there is an element $u \in G$ such that*

$$uau^{-1}=b. \tag{9.32}$$

Note the following *properties of* conjugate elements.

(1) *Every element a is the conjugate of itself.* Indeed, if e is the unity of a group, then, evidently, the relation $eae^{-1}=a$ holds true, which means that a is the conjugate element of a.

(2) *If the element b is the conjugate of a, then the element a is the conjugate of b.* This property follows immediately from (9.32). Indeed, multiplying both sides of (9.32) by u^{-1} on the left and by u on the right, we find that $u^{-1}bu=a$. Noting that the inverse element of u^{-1} is u, we verify the validity of the property we have formulated.

(3) *If b is the conjugate element of a and c is the conjugate element of b, then c is the conjugate of a.*

Indeed, since $c=vbv^{-1}$, and $b=uau^{-1}$, then, evidently,

$$c=vuau^{-1}v^{-1}. \tag{9.33}$$

Since the inverse element of vu is $u^{-1}v^{-1}$, it follows from (9.33) that the element c is the conjugate of a.

Let us place in the same class all the elements of the group which are the conjugates of the given element a. Thus, according to property (3) every element of the class is the conjugate of any element of that class. Two classes of this kind evidently either coincide or have no elements in common.

Let us return now to group representations.

Assume that a and b are conjugate elements, i.e. relation (9.32)

holds true:

$$b = uau^{-1}. \tag{9.32}$$

Let us direct our attention to the operators $D(a)$, $D(b)$, $D(u)$ and $D(u^{-1})$. According to the definition of a group representation, the operator $D(u^{-1})$ is an inverse of the operator $D(u)$, i.e. $D(u^{-1}) = (D(u))^{-1}$.

Using again the definition of a representation, we get, in accordance with (9.32), a relation $D(b) = D(u)D(a)(D(u))^{-1}$.

Let us pass now to the matrices of the operators appearing in the last relation. We see that we can regard the matrix of the operator $D(b)$ as that of the operator $D(a)$ when we pass to a new basis with the transition matrix $D(u)$ (see 5.2.2). Since under such transformations the trace of the matrix is invariant and by definition is equal to the character of the element, we can infer that $\chi(a) = \chi(b)$.

Thus *the characters of all elements belonging to the same class of conjugate elements are equal to each other.*

It is also evident that *the characters of the elements for equivalent representations coincide.*

The concept of a character is usually used in the theory of representations as follows.

Assume that the given group G can be divided into a finite number of various classes of the conjugate elements $K_1, K_2, \ldots, K_\nu$. Then every element of the class K_i in the given representation $D(G)$ (and in any equivalent representation) is associated with the same character χ_i. Therefore, the representation $D(G)$ can be described by means of a collection of characters $\chi_1, \chi_2, \ldots, \chi_\nu$ which can be regarded as the coordinates of a vector in a Euclidean space of dimension ν. Thus different vectors correspond to different representations. The indicated geometrical approach makes it possible to solve important problems of the group representation theory in many cases.

9.3.5. Examples of group representations.

Example 1. Assume that G is a symmetry group of a three-dimensional space consisting of two elements: of an identity transformation $\boldsymbol{I}$ (the unity of the group) and the reflection $\boldsymbol{P}$ about the origin. Thus $G = \{\boldsymbol{I}, \boldsymbol{P}\}$.

The multiplication of the elements of the group is defined by the following table:

	I	P
I	I	P
P	P	I

(9.34)

(1) *A one-dimensional representation of the group G.*

We choose a basis $\boldsymbol{e}_1$ in the space E^1 and consider the matrix $A^{(1)}$ of the linear nonsingular transformation $\boldsymbol{A}^{(1)}$ in that space: $A^{(1)}=(1)$. The transformation $\boldsymbol{A}^{(1)}$ evidently forms a subgroup in the group $GL(1)$ of the linear transformations of the space E^1, the multiplication in this subgroup being defined by the table

	$A^{(1)}$
$A^{(1)}$	$A^{(1)}$

We evidently get a one-dimensional representation $D^{(1)}(G)$ of the group G by means of the relations $D^{(1)}(\boldsymbol{I})=\boldsymbol{A}^{(1)}$, $D^{(1)}(\boldsymbol{P})=\boldsymbol{A}^{(1)}$ (these relations define the homomorphism of the group G into $GL(1)$ and, consequently, its representation).

(2) *A two-dimensional representation of the group G.*

We choose some basis $\boldsymbol{e}_1$, $\boldsymbol{e}_2$ in E^2 and consider in that basis the matrices $A^{(2)}$ and $B^{(2)}$ of linear nonsingular transformations $\boldsymbol{A}^{(2)}$ and $\boldsymbol{B}^{(2)}$:

$$A^2=\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad B^2=\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

(since $\det \boldsymbol{A}^{(2)}=1$ and $\det \boldsymbol{B}^{(2)}=-1$, we find that $\boldsymbol{A}^{(2)}$ and $\boldsymbol{B}^{(2)}$ are nonsingular transformations).

The transformations $\boldsymbol{A}^{(2)}$ and $\boldsymbol{B}^{(2)}$ form a subgroup in the group $GL(2)$. We make sure by a direct verification (by multiplying the matrices $A^{(2)}$ and $B^{(2)}$) that the multiplication of the operators $\boldsymbol{A}^{(2)}$ and $\boldsymbol{B}^{(2)}$ is defined by the table

	$A^{(2)}$	$B^{(2)}$
$A^{(2)}$	$A^{(2)}$	$B^{(2)}$
$B^{(2)}$	$B^{(2)}$	$A^{(2)}$

(9.35)

We obtain a two-dimensional representation $D^{(2)}(G)$ of the group G by the relations

$$D^{(2)}(\boldsymbol{I})=\boldsymbol{A}^{(2)}, \quad D^{(2)}(\boldsymbol{P})=\boldsymbol{B}^{(2)}. \tag{9.36}$$

Indeed, comparing tables (9.34) and (9.35), we see that (9.36) defines the isomorphism of the group G into the subgroup $\{\boldsymbol{A}^{(2)}, \boldsymbol{B}^{(2)}\}$ of the group $GL(2)$ and, consequently, the representation of this group.

(3) *A three-dimensional representation of the group G.*

Let us consider in E^3 a linear representation $\boldsymbol{A}^{(3)}$ defined by the matrix

$$\boldsymbol{A}^{(3)}=\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

This transformation forms a subgroup of the group $GL(3)$ with the multiplication law $\boldsymbol{A}^{(3)}\boldsymbol{A}^{(3)}=\boldsymbol{A}^{(3)}$.

As in the case of a one-dimensional representation, we get a three-dimensional representation $D^{(3)}G$ by means of the relations

$$D^{(3)}(\boldsymbol{I})=\boldsymbol{A}^{(3)},\ D^{(3)}(\boldsymbol{P})=\boldsymbol{A}^{(3)}.$$

(4) *A four-dimensional representation of the group G.*

Let us consider in E^4 the linear transformations $\boldsymbol{A}^{(4)}$ and $\boldsymbol{B}^{(4)}$ defined by the matrices

$$A^{(4)}=\left(\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right), \qquad B^{(4)}=\left(\begin{array}{cc|cc} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{array}\right).$$

The transformations $\boldsymbol{A}^{(4)}$ and $\boldsymbol{B}^{(4)}$ form a subgroup in the group $GL(4)$, with the multiplication law defined by a table similar to (9.35) (with the index 2 replaced by 4). We evidently obtain a four-dimensional representation $D^{(4)}(G)$ of the group G by the relations

$$D^{(4)}(\boldsymbol{I})=\boldsymbol{A}^{(4)},\ D^{(4)}(\boldsymbol{P})=\boldsymbol{B}^{(4)}.$$

Remark. It is easy to see that the matrices $A^{(4)}$ and $B^{(4)}$ can be written in the form

$$A^{(4)}=\begin{pmatrix} A^{(2)} & 0 \\ 0 & A^{(2)} \end{pmatrix}, \quad B^{(4)}=\begin{pmatrix} B^{(2)} & 0 \\ 0 & B^{(2)} \end{pmatrix}.$$

The representation $D^{(4)}(G)$ can, therefore, be conventionally written in the form $D^{(4)}(G)=D^{(2)}(G)+D^{(2)}(G)=2D^{(2)}(G)$. Quite similarly we can conventionally write $D^{(3)}(G)$ in the form $D^{(3)}(G)=3D^{(1)}(G)$.

Using this remark, the reader will easily construct a representation of the group G of any finite dimensionality.

Example 2. We have proved in 9.2.5 that the symmetry group $G=\{\boldsymbol{I}, \boldsymbol{P}\}$ of a three-dimensional space just considered was a normal subgroup of the group $O(3)$ (a group of orthogonal transformations of the space E^3). We proved there that the subgroup $SO(3)$ of the proper orthogonal transformations of the group $O(3)$ was isomorphic to the factor group of the group $O(3)$ by the normal subgroup $\{\boldsymbol{I}, \boldsymbol{P}\}$.

Since the group is homomorphically mapped onto each of its factor groups, it follows that $O(3)$ can be homomorphically mapped onto the group $SO(3)$. As we saw in 9.3.5, the indicated homomorphism was carried out as follows.

If a is a proper transformation from $O(3)$, it is associated with the same transformation from $SO(3)$.

If a' is an improper transformation, then it is associated with a proper transformation Pa'.

We thus obtain a three-dimensional representation $DO(3)$ of the group of orthogonal transformations by means of the group $SO(3)$ of proper orthogonal transformations.

TO THE READER

Mir Publishers welcomes your comments on the content, translation and design of this book.

We would also be pleased to receive any other suggestions you may wish to make.

Our address is:
Mir Publishers
2, Pervy Rizhsky Pereulok
I-110, GSP, Moscow, 129820
USSR

Printed in the Union of Soviet Socialist Republics

ANALYTIC GEOMETRY

V. Ilyin, D.Sc. (Phys.-Math.), and E. Poznyak, D.Sc. (Phys.-Math.)

This book is intended for university students majoring in various fields of physics and applied mathematics. It arises from a course of lectures in analytic geometry, delivered by the authors at the Physics Faculty of Moscow State University.
The text consists of seven chapters and an appendix. Each chapter is supplied with a large number of worked problems which are regarded as an integral part of the book.
Translated from the third (1981) Russian edition.

LIE GROUPS AND LIE ALGEBRAS

M. Postnikov, D.Sc. (Phys.-Math.)

This textbook like the rest of the series is compiled from lectures delivered by the author to students at Moscow State University. The audience was made up of final year students and postgraduates, and, as a consequence, a certain level of knowledge (for example, the properties of manifolds) was assumed.

The lectures fall into five parts. The first covers the basic concepts of Lie groups, Lie algebras, and the Lie algebra of a given Lie group. The second part deals with "locality theory", the third generalizing these ideas. The first three sections in fact could be used as a foundation course in Lie algebra for beginners. The fourth part considers Lie subgroups and Lie factor groups. The final part has a purely algebraic character and is practically independent of the preceding four parts; formally it considers a proof of Ado's theorem, which is interesting in itself.

This book will be useful to mathematics students taking courses in Lie algebra because it contains all the major results and proofs in a compact and accessible form.